Hürth, den 24. März 2020

Universität zu Köln
Sprachliche Informationsverarbeitung
Themensteller: Vertr.-Prof. Dr. Nils Reiter

Concept Mining – Disambiguierung extrahierter Terme am Beispiel von Stellenausschreibungen

vorgelegt von

Johanna Binnewitt
Matrikelnummer 5429692
E-Mail: binnewij@smail.uni-koeln.de

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

Digitale Technologien ermöglichen heute in verschiedenen Bereichen maschinelle Datenverarbeitungen, die manuell nicht zu bewältigen sind. Unter dem Sammelbegriff Data Mining finden sich verschiedenste Unterdisziplinen zusammen, die beispielsweise Datenmengen auf Muster hin untersuchen. Methoden, die dabei Text als Datengrundlage nutzen, fasst man auch mit dem Begriff Text Mining zusammen. Dabei wird unstrukturierter Text wird verwendet, um darin geäußerte Meinungen zu klassifizieren, neu aufkommende Themen zu entdecken oder die Informationen über Ereignisse aus dem Text zu extrahieren (Prasad et al. 2017, S. 7). Auch bei Analysen, die den Arbeitsmarkt betreffen, spielt Text Mining eine immer wichtigere Rolle. Als Datengrundlage dienen dabei zum Beispiel Lebensläufe von Bewerber*innen oder die Texte von Stellenausschreibungen, mit deren Hilfe Erkenntnisse über Bedarf und Nachfrage auf dem Arbeitsmarkt gewonnen werden sollen. In den Stellenanzeigen sind dafür vor allem Passagen interessant, in denen berichtet wird, welche Kompetenzen der zukünftige Arbeitnehmer mitbringen soll. Das Projekt *Qualifikationsentwicklungsforschung* beschäftigt sich seit 2015 damit, wie Methoden aus dem Bereich des Text Minings dazu verwendet werden können, diese Informationen aus Stellenanzeigen zu extrahieren. Dabei liegt eine Herausforderung darin, Text mit semantischem Inhalt zu füllen. Denn nur so gelingt eine tatsächliche Verknüpfung zwischen der als Text vorliegenden Stellenanzeige und den darin angesprochenen Qualifikationen und Fähigkeiten. Eine Extraktion von Bedeutung, die nicht nur auf die Identifizierung von Schlüsselwörtern aufbaut, wird auch als *Concept Mining* bezeichnet. Hier wird die Bedeutung von Ausdrücken und nicht das Wort selbst in den Mittelpunkt gestellt (Prasad et al. 2017, S. 7). Dieser Schritt ist wichtig, weil Sprache variabel ist und sich deshalb verschiedenste Formulierungen auf ein und dieselbe Kompetenz beziehen können. Dabei ist allerdings nicht immer offensichtlich, welcher Ausdruck an welches Konzept gebunden ist, was auch an der Ambiguität von Sprache liegt. Ambiguität bedeutet zunächst einmal, dass ein sprachlicher Ausdruck verschiedene Bedeutungen haben kann. Die Identifizierung der jeweils aktivierten Bedeutung ist für uns Menschen trivial. Der Kontext bietet uns meist genügend Informationen, um jeweils zu wissen, was gemeint ist. Jedoch stellt die Zuweisung einer Bedeutung die Softwareentwicklung vor eine Herausforderung. In der Computerlinguistik beschäftigen sich Forscher*innen im Bereich der *Word Sense Disambiguation* (WSD) damit, computergestützt zu entscheiden, welche Bedeutung bei einem Wort in einem spezifischen Kontext aktiviert wird (Agirre und Edmonds 2007, S. 1–2). Eine solche Entscheidung würde es ermöglichen, Stellenanzeigen mit

Bedeutungen anzureichern. Im Gegensatz zu uns Menschen muss in einem maschinellen Verfahren explizit implementiert werden, wie die Verknüpfung von Ausdruck und Inhalt ablaufen soll (Agirre und Edmonds 2007, S. 1). Diese Herausforderung wird in dieser Arbeit am Beispiel von Stellenausschreibungen näher beleuchtet. Dabei spielen einerseits mehrdeutige Wörter eine Rolle, wie beispielsweise in der Formulierung *Erfahrung in der Administration von Entwicklungsumgebungen*. Andererseits kommt die semantische Ebene zum Tragen, um eine Unterscheidung der Bedeutung des Ausdrucks *französisch* zwischen den Sätzen *Wir betreuen französisch Unternehmen* und *Sie benötigen französische Sprachkenntnisse* vornehmen zu können.

Kapitel 2 beginnt mit allgemeinen Überlegungen zur Beziehung zwischen Form und Inhalt. Diese Beziehung kann durch verschiedene Relationen ausgedrückt werden, die vorgestellt werden sollen. Nach einem kurzen Überblick zu den existierenden Verfahren zur WSD sowie zu Text Mining-Projekten im Bereich der Arbeitsmarktforschung beginnt in Kapitel 4 der Einstieg in das Forschungsgebiet der Stellenausschreibungen. Hier werden einerseits das Projekt *Qualifikationsentwicklungsforschung* und andererseits existierende Kompetenz-Taxonomien, die Qualifikationen und Fähigkeiten systematisch zu einander in Beziehung setzen, vorgestellt. Außerdem werden erste explorative Methoden umgesetzt, die zeigen sollen, welche sprachlichen Eigenschaften auf Stellenanzeigen zutreffen. Kapitel 5 beschäftigt sich konkret mit der Umsetzung von Disambiguierung am Beispiel von Stellenanzeigen. Dafür wird ein Ansatz erprobt und anschließend evaluiert. In Kapitel 6 geschieht schließlich dasselbe für die Identifizierung von Synonymen, die vor allem eine Rolle bei neu identifizierten Ausdrücken und ihrer Einordnung in die Taxonomie spielt.

2. Beziehung zwischen Form und Inhalt

Versucht man mithilfe von Stellenanzeigen die Anforderungsprofile von Berufen zu analysieren, stößt man schnell auf ein Problem, das sich generell bei der Informationsgewinnung aus Texten ergibt: Die als Text kodierte Information muss dekodiert werden, um die dahinterliegende Bedeutung zu entschlüsseln. Was für Menschen eine alltägliche Tätigkeit darstellt, ist für ein maschinelles System zunächst eine Herausforderung. Bei der Analyse von Bedeutung in Texten soll von einem String, der unstrukturierten Text enthält, auf die dahinterliegende Bedeutung geschlossen werden. In Bezug auf Stellenanzeigen kann beispielsweise ein*e Bewerber*in beim Lesen einer Anzeige erkennen, welche Kompetenzen in der ausgeschriebenen Stelle gefordert werden und ob dieses Profil auf sie bzw. ihn zutrifft. Ein maschinelles System jedoch muss zum einen „wissen“, wo die relevante Information, also die Kompetenz, im Text genannt ist, und zum anderen erkennen, um welche Art von Kompetenz es sich handelt. Der erste Schritt kann mithilfe von klassischen Methoden aus dem Gebiet der Informationsextraktion (IE) bearbeitet werden, die dazu dienen, relevante Informationen in einem unstrukturierten Text zu identifizieren (Cowie und Lehnert 1996). Anwendungen aus diesem Bereich helfen dabei relevante Textfragmente zu erkennen und aus den Fragmenten die gesuchte Information zu extrahieren. Dabei werden die Informationsbausteine wieder in einer anderen Struktur, zum Beispiel in einer Datenbank, aufbereitet (Cowie und Lehnert 1996, S. 81). Eine Anwendung, die IE für Stellenanzeigen betreibt, liegt bereits vor (Geduldig 2017). Daher ist IE selbst nicht Gegenstand der vorliegenden Arbeit. Vielmehr baut die in dieser Arbeit präsentierte Anwendung auf dem IE-System von Geduldig auf. Mithilfe dieses Frameworks können Nutzer*innen automatisiert relevante Zeichenketten in Stellenausschreibungen identifizieren. Anhand welcher Methoden dies geschieht wird in Kapitel 3.1 näher erläutert.

Wenn man sich die Motivation hinter IE jedoch näher anschaut, steht häufig bei der Extraktion von Informationen nicht die Extraktion von Strings im Mittelpunkt. Vielmehr verbirgt sich dahinter der Wunsch, die durch den String repräsentierte Bedeutung zu extrahieren. Dabei gibt es verschiedene Hindernisse: Zum einen können Stellenanzeigen anaphorische Ausdrücke beinhalten. Damit sind zunächst Formulierungen gemeint, bei denen sich ein Ausdruck nur in Abhängigkeit zum vorangegangenen Kontext deuten lässt (Rösiger 2019, S. 22). Als Beispiel dafür betrachten wir folgenden Satz:

> Arminia Bielefeld spielt wieder oben mit. Sie gewannen am Wochenende gegen Greuther Fürth.

Das Pronomen zu Beginn des zweiten Satzes erhält seine Bedeutung nur durch die Wörter aus dem vorherigen Satz. Ein ideales IE-System sollte im genannten Beispiel in der Lage sein, die Anapher zwischen „Arminia Bielefeld“ und „Sie“ zu erkennen und entsprechend aufzulösen. Dadurch wäre es dem IE-System möglich zu extrahieren, dass Arminia Bielefeld das Spiel gewonnen hat. Neben der Verwendung von Pronomen können Umschreibungen ebenfalls Koreferenz erzeugen. Zum Beispiel, wenn man die obige Anapher umformuliert:

> Arminia Bielefeld spiel wieder oben mit. Die Kicker Ostwestfalens gewannen am Wochenende gegen Greuther Fürth.

Die Kicker Ostwestfalens referenziert hier ebenfalls auf den Verein Arminia Bielefeld. Diese Verbindung kann allerdings nur auf Grundlage von Wissen hergestellt werden, das außerhalb des Textes vorliegt. In Ansätzen wird die Auflösung koreferenzieller Ausdrücke im Rahmen von IE-Systemen bereits bearbeitet. Dabei gibt es vor allem Fortschritte bei der Bearbeitung von englischsprachigen Texten. So entwickelten Clark und Manning eine Anwendung, die mithilfe eines Künstlichen Neuronalen Netztes entscheidet, ab wann zwei in einem Text genannte Entitäten zu einer verschmelzen (Clark und Manning 2016). Doch auch im Rahmen der Behandlung von deutscher Sprache gibt es Entwicklungen. Dafür sollen unter anderem funktionierende Ansätze zur Koreferenz-Auflösung aus dem Englischen ins Deutsche adaptiert werden (Rösiger 2019, S. 83).

Ansätze aus dem Bereich des Concept Minings gehen im Vergleich zu Koreferenz-Auflösungen allgemeiner vor, indem sie nicht nur für koreferenzielle Formulierungen die Bedeutungen, die im Text auftauchen, identifizieren und extrahieren. Stattdessen wird für jeden als relevant eingestuften String im Text eine Verknüpfung zum dahinterliegenden semantischen Konzept (siehe 2.1) erzeugt. Dadurch soll zusätzlich zur Auflösung der Koreferenz die korrekte Auflösung von ambigen Ausdrücken (siehe 2.2.) gewährleistet werden. Concept Mining-Ansätze können zum Beispiel Taxonomien verwenden, wie sie in Kapitel 3.2 näher beschrieben werden. Taxonomien bilden Klassen auf „der Grundlage von Ähnlichkeitsbeziehungen zwischen den der Taxonomie unterworfenen Gegenständen“, also den Kompetenzen (Metzler Lexikon Sprache, 5. Auflage, s.v. „Taxonomie“). Jedoch ist die Zuordnung zwischen einem Ausdruck im Text und einem Eintrag in der Taxonomie nicht immer trivial. Deshalb wird nun erst näher dargestellt, welche Beziehungen zwischen Ausdrücken und Inhalten bestehen können. Dafür werden zunächst einige notwenige Begriffe aus der Zeichentheorie näher definiert.

Ausgehend von Ferdinand de Saussure wird ein sprachliches Zeichen in der Linguistik als ein Zusammenschluss von Inhalt und Ausdruck definiert. Letzteres, auch Signifikant genannt, „ist dessen Lautgestalt, d.h. eine Einheit, die eine Sequenz von Lauten darstellt“ (Schwarz und Chur 2007, S. 22). Sobald einer solchen Lautgestalt ein Inhalt zugeordnet ist, handelt es sich um ein sprachliches Zeichen. Inhalt (auch Signifikat) meint dabei, dass das Zeichen „Informationen über bestimmte Gegenstandsbereiche“ transportiert. Wenn Inhalt und Ausdruck in einer Sprachgemeinschaft unzertrennbar verbunden sind, spricht man auch von einer bilateralen Zeichenkonzeption (Schwarz und Chur 2007, S. 22). Der Ausdruck *Schlüssel* aktiviert beispielsweise direkt eine bestimmte Vorstellung in unserem Kopf. Diese ist abhängig von der Sprachgemeinschaft, zum Beispiel ist die Lautgestalt *Schlüssel* im Ungarischen mit keinem Inhalt verknüpft. Die Aktivierung des Inhalts ist also sprachgebunden. Rothe und Schütze bezeichnen diese Vereinigung von einer Schreibweise bzw. Aussprache und einem Inhalt auch als Lexem (Rothe und Schütze 2015, S. 1793). In dieser bilateralen Beziehung fehlt allerdings noch der Bezug zur Welt. Die Gegenstände in der Welt, die durch sprachliche Zeichen benannt werden, werden Referenten genannt (Schwarz und Chur 2007, S. 22). Mithilfe des Inhalts, der an einen Ausdruck geknüpft ist, kann eine Referenz in die Welt, wie beispielsweise auf einen real existierenden Fahrradschlüssel, erzeugt werden. Diese Beziehung wird auch häufig im semiotischen Dreieck dargestellt (siehe Abbildung 1), bei dem das Symbol, also der Ausdruck, einen Begriff, also den Inhalt, erweckt. Beide stehen für bzw. beziehen sich auf einen Gegenstand, der in der realen Welt existiert.

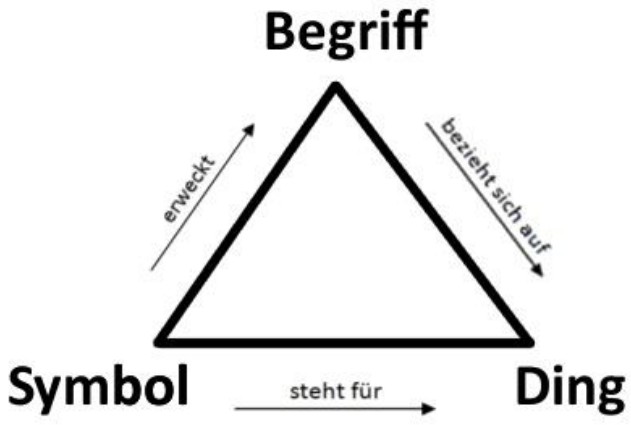

Abbildung 1: Semiotisches Dreieck (Quelle: https://commons.wikimedia.org/wiki/File:Semiotischesdreieck.jpg)

Die meisten Zeichen basieren auf einer arbiträren, also willkürlichen, Beziehung, die oft auf Konventionen beruht. Als Außenstehender ist es also meist nicht möglich von einem

sprachlichen Zeichen auf den verknüpften Inhalt zu schließen (Schwarz und Chur 2007, S. 23). Offenbar sind die beiden Teile jedoch innerhalb einer Sprachgemeinschaft fest verknüpft. Im Folgenden wollen wir uns die Inhaltsseite eines sprachlichen Zeichens näher anschauen, da diese eine zentrale Rolle für das Concept Mining spielt.

2.1. Das semantische Konzept

Die lexikalische Semantik beschäftigt sich mit der Inhaltsseite von sprachlichen Zeichen auf Wortebene. Dabei spielen sogenannte Konzepte eine wichtige Rolle. Schwarz und Chur definieren diese wie folgt:

> „Konzepte (auch: Begriffe) sind die Bausteine unseres Wissens. Konzepte sind mentale Einheiten; sie basieren auf Erfahrungen, die wir im Umgang mit der Welt machen. Im [Langzeitgedächtnis] haben wir kategoriales und individuell-episodisches Wissen über die Welt gespeichert. Kategoriales Wissen ist allgemeines Wissen über die Welt, Wissen über Klassen von Gegenständen. Einheiten, die Informationen über ganze Klassen repräsentieren, sind Kategorien(konzepte).“ (Schwarz und Chur 2007, S. 24)

Demnach gibt es beispielsweise ein Kategorienkonzept *Schlüssel*, welches alle Gegenstände, die der Klasse *Schlüssel* angehören, zu einer gemeinsamen mentalen Einheit zusammenfasst. Bei dem Satz *Der Schlüssel passt ins Schloss*. haben wird diese mentale Einheit aktiviert, sodass wir eine Vorstellung davon bekommen, was mit dem Ausdruck gemeint ist. Den Kategorienkonzepten stehen die Partikularkonzepte gegenüber. Diese speichern Informationen über Entitäten, wie beispielsweise bestimmte Gegenstände, Situationen oder Personen. Dieses Wissen ist an räumliche und zeitliche Erfahrungen gebunden, weshalb es auch individuell-episodisches Wissen genannt wird (Schwarz und Chur 2007, S. 25). Der Fahrradschlüssel aus dem vorherigen Beispiel wird beispielsweise durch ein Partikularkonzept im Gehirn des Besitzers abgebildet, das unter anderem Informationen zu dessen spezifischen Farbe oder Form speichert. Gleichzeitig wird dieser Fahrradschlüssel aber auch mit allem Wissen über das Kategorienkonzept *Schlüssel* verknüpft. Dies geschieht mithilfe des Prinzips der Äquivalenz. Es „ermöglicht die Klassifizierung von zwei verschiedenen Objekten […] als Vertreter (Instanzen) einer Klasse“ (Schwarz und Chur 2007, S. 25). Dadurch wird uns vor allem ermöglicht, Entitäten der Welt in Kategorien zu ordnen und uns über diese Entitäten zu verständigen.

2.1.1. Äquivalenz in der Prototypentheorie

Wie Äquivalenz zustande kommt, wird durch zwei verschiedene Modelle erklärt: die Prototypentheorie und die Merkmalshypothese. Laut der Prototypentheorie wird ein

Kategorienkonzept von einer Instanz der Klasse repräsentiert, dem Prototypen. Für diese Rolle kommt zunächst jede Entität, die zur Kategorie gehört, in Frage, jedoch verfügen diese im Einzelnen über ein unterschiedliches Maß an Repräsentativität. Deshalb wird der Prototyp so gewählt, dass er möglichst typisch für die jeweilige Kategorie ist (Schwarz und Chur 2007, S. 49). Beispielsweise wird die Kategorie *Schlüssel* vermutlich häufig durch eine mentale Einheit repräsentiert, die aus Metall ist und Zacken hat. Diese Eigenschaft trifft auf viele Entitäten zu, die der Kategorie angehören. Gleichzeitig gehört ein elektronischer Autoschlüssel dem Konzept genauso an. Dieser wird aber wahrscheinlich nur in den seltensten Fällen als Prototyp gewählt, weil er die Menge aller Schlüssel nicht gut genug repräsentiert. Die Verwendung von Prototypen hat laut Schwarz und Chur vor allem auch ökonomische Gründe. Durch sie ist es uns möglich, den Speicherplatz unseres mentalen Lexikons optimal zu nutzen. Sobald wir eine neue Entität kennenlernen, können wir sie einem uns schon bekannten Prototypen anhängen, ohne dass wir die gesamte Kategorie neu umstrukturieren müssen (Schwarz und Chur 2007, S. 49). Dieses Prinzip gilt sowohl für materielle Konzepte wie beispielsweise Gegenstände als auch für abstraktere Gebilde wie „kategoriale Zustands-, Vorgangs- und Handlungskonzepte“ (Schwarz und Chur 2007, S. 51). Abstrakte Konzepte sind vor allem für die Kategorisierung von Kompetenzen in Stellenanzeigen interessant. Zum Beispiel ist die Formulierung „Textverarbeitungsprogramme verwenden“ mit einem Handlungskonzept verknüpft, das mental verankert ist und das laut der kognitiven Linguistik bei jedem unterschiedlich ausgeprägt sein kann. Unter diesem Konzept sammeln wir verschiedene Handlungen wie beispielsweise „das Layout für einen Text festlegen“ oder „Texte mit einem Textverarbeitungsprogramm schreiben“. Wenn man nun gerade dabei ist einen Text mit einem Textverarbeitungsprogramm zu verfassen, wird diese Tätigkeit auf Grundlage der Äquivalenz mit dem Handlungskonzept verknüpft. Also vereint dieses Handlungskonzept verschiedene spezifische Handlungen, genau wie ein materielles Kategorienkonzept unterschiedliche Partikularkonzepte bündelt. Die Prototypentheorie behandelt außerdem die Frage, wie Konzepte untereinander in Beziehung stehen. Dabei wird davon ausgegangen, dass die „Einheiten unserer Erfahrungswelt […] in Taxonomien klassifiziert“ werden. Wir bilden also Hierarchien, um die uns bekannten Kategorienkonzepte miteinander in Beziehung zu setzen. Die Ebene, auf der zwischen den Kategorien die größte Abgrenzung stattfindet, nennt man auch Basisebene (Schwarz und Chur 2007, S. 52). Beispielsweise grenzen sich die Konzepte für *Blume* und *Stuhl* stärker voneinander ab als die von *Tulpe* und *Narzisse*. Während alle vier Konzepte von Prototypen repräsentiert werden, stimmen die beiden Prototypen der letzten beiden Kategorienkonzepte stärker überein als die von *Blume* und *Stuhl*, weil sich Tulpen und Narzissen einige Eigenschaften wie zum

Beispiel grüne Blätter und Blüten teilen. In Bezug auf die hierarchische Strukturierung von Konzepten kann es passieren, dass das übergeordnete Konzept keinen klar umrissenen Prototypen hat (Cruse 2000, S. 37). Beispielsweise wird das übergeordnete Konzept von *Stuhl* und *Sessel* durch keinen klaren Prototyp dargestellt, der sich von den spezifischen Konzepten abgrenzt. Das Konzept *Sitzgelegenheit* wird eher durch einen Vertreter der Unterkonzepte, wie beispielsweise einen Stuhl, repräsentiert.

2.1.2. Äquivalenz durch die Merkmalshypothese

Neben der Prototypentheorie versucht außerdem die Merkmalshypothese Äquivalenz zu erklären. Bei dieser wird davon ausgegangen, dass sich Bedeutungen aus kleineren Einheiten, den semantischen Merkmalen, zusammensetzen (Schwarz und Chur 2007, S. 37). Auf dieser Grundlage kann zum einen entschieden werden, ob eine Entität einem Kategorienkonzept angehört, indem alle für die Kategorie typischen Merkmale mit den Merkmalen der Entität verglichen werden. Ist die Übereinstimmung groß genug, so gehört die Entität der ausgewählten Kategorie an. Zum anderen ist es durch die Merkmalshypothese nun auch einfacher Konzepte untereinander zu vergleichen. In diesem Zusammenhang sprechen Turney und Pantel von attributionaler Ähnlichkeit zwischen Wörtern und meinen damit, dass sich hier die Merkmale der Konzepte, die diese Begriffe repräsentieren, überschneiden (Turney und Pantel 2010, S. 149). Beispielsweise teilen sich die Konzepte *Sofa* und *Sessel* die Merkmale, dass sie Sitzgelegenheiten sind und vorwiegend aus Polstern bestehen. Gleichzeitig dienen Merkmale dazu, Bedeutungen voneinander abzugrenzen. Indem man beispielsweise die distinktiven Merkmale von *Sofa* und *Sessel*, zum Beispiel, die Anzahl der Sitze pro Möbelstück, gegenüberstellt, erhält man semantische Oppositionen (Schwarz und Chur 2007, S. 37). Sobald andererseits die Merkmale von zwei Ausdrücken vollständig übereinstimmen, handelt es sich um Synonyme (siehe 2.3.). Zwar lassen sich viele Gegenstände und mentale Konzepte auf den ersten Blick durch Merkmale beschreiben und voneinander unterscheiden, jedoch stößt die Merkmalshypothese im Hinblick auf die Granularität von Bedeutung schnell an ihre Grenzen. Deshalb wird sie an vielen Stellen durch die zuvor beschriebene Prototypentheorie ergänzt, da Bedeutungen „in ihren Relationen nicht immer binär beschreibbar, sondern graduell“ sind (Schwarz und Chur 2007, S. 53). Diese Eigenschaft kann durch prototypische Vertreter einer Bedeutungskategorie viel besser abgebildet werden.

2.1.3. Semantische Konzepte und Bedeutung

Sobald ein Konzept durch eine sprachliche Form ausgedrückt wird, ist dem Konzept eine Bedeutung zugeordnet (Schwarz und Chur 2007, S. 26). Dies trifft nicht auf alle Konzepte zu,

was dadurch veranschaulicht werden kann, dass es in jeder Sprache Wörter gibt, die nicht in jede andere beliebige Sprache übersetzt werden können. Beispielsweise lässt sich das Wort „Fernweh“ nur bedingt in andere Sprachen übersetzen. Trotzdem gibt es in anderen Sprachgemeinschaften eine mentale Einheit, die zusammenfasst, dass man sich nach fernen Orten sehnt. Hier werden dann Umschreibungen gebildet, wie zum Beispiel „envie de voyager“ (Reiselust) im Französischen, die dem Konzept möglichst nahe kommen sollen. Da diese Arbeit ein Korpus als Grundlage verwendet, interessieren uns hier vorwiegend die versprachlichten Konzepte, die eine Bedeutung haben. Bedeutung ist vor allem kontextabhängig. Das zeigt sich bereits durch die Sprachabhängigkeit von Zeichen, da die gleiche Lautgestalt in zwei verschiedenen (Sprach-)Kontexten vollkommen unterschiedliche Inhalte aktivieren kann. Beispielsweise bedeutet die Zeichenkette *gift* im Englischen „Geschenk“, während sie im Deutschen für eine toxische Substanz steht. Je nach sprachlichem Kontext kann der Ausdruck also vollkommen andere mentale Konzepte aktivieren. Der sprachliche Kontext teilt sich in einen soziokulturellen Umstand, in dem ein Ausdruck geäußert wird und einen rein linguistischen Kontext, der aus den unmittelbar umliegenden Wörtern besteht (Sahlgren 2006, S. 26). Für diese Arbeit wird der soziokulturelle Kontext in den Überlegungen vernachlässigt. Zum einen ist dieser vermutlich in allen Texten ähnlich, da man Stellenanzeigen an sich schon als soziokulturellen Umstand auffassen kann. Außerdem ist es schwierig die Messung eines solchen Kontexts rein auf Grundlage von Texten zu operationalisieren. Stattdessen werden wir uns der Untersuchung des linguistischen Kontexts widmen. Die einzige Überlegung an dieser Stelle ist, ob sich der soziolinguistische Kontext nicht gleichzeitig auch im rein linguistischen Kontext manifestiert. Die Änderung im Sprachgebrauch müsste sich auch im unmittelbaren Kontext, beispielsweise durch die anderen Wörter im Satz, bemerkbar machen. Denn die unterschiedlichen Bedeutungen in verschiedenen soziolinguistischen Verhältnissen müssen sich in gewisser Form auch auf den Kontext auswirken. Durch die Beziehung zwischen Ausdruck, Kontext und Bedeutung ergibt sich zwangsweise, dass sprachliche Zeichen nicht immer eine 1:1-Beziehung zwischen Form und Inhalt darstellen. In dieser Arbeit sind vor allem Ambiguität und Synonymie zwei wichtige semantische Relationen, die deshalb im Folgenden vorgestellt werden.

2.2. Variabilität und Ambiguität

Wenn der Kontext die Bedeutung eines Ausdrucks beeinflusst, ist der Umkehrschluss, dass auf jede sprachliche Äußerung isoliert betrachtetet mehrere Bedeutungen zutreffen können. Dieses Phänomen wird auch als Variabilität bezeichnet und entsteht entweder in einer diachronen Perspektive – in einem Sprachwandel über die Zeit hinweg – oder in einer synchronen

Perspektive, wenn gleiche sprachliche Ausdrücke mit unterschiedlichen Bedeutungen besetzt werden (Neuefeind 2019, S. 9). Die Variabilität von Sprache macht sich zum einen in der semantischen Vagheit von Begriffen bemerkbar. Damit ist die „interpretatorische Unbestimmtheit hinsichtlich einiger weniger semantischer Merkmale bei einer festen Kernbedeutung“ gemeint. Dieses Phänomen tritt beispielsweise bei Adjektiven wie *klein* oder *groß* auf, bei denen neben einer Kernbedeutung immer subjektive Auffassungen mit einfließen können (Neuefeind 2019, S. 10). Sobald allerdings zu viele der für die Kernbedeutung wichtigen Merkmale voneinander abweichen, spricht man von einer Ambiguität. Ambiguität tritt immer dann auf, wenn mehrere Bedeutungen zu einer Wortform zugeordnet sind. In der Lexikografie wird diese semantische Relation oft so dargestellt, dass unter einem Lexikoneintrag mehrere Lexeme eingetragen sind. Bei Ambiguität unterscheidet man zwischen Polysemie und Homonymie. Ersteres meint, dass alle Bedeutungen eines Ausdrucks auf eine gemeinsame Kernbedeutung zurückgeführt werden können. Administration im Sinne der Serveradministration hat beispielsweise eine andere Bedeutung als Administration im Sinne der Personaladministration. Trotzdem liegt beiden Bedeutungen die Kernbedeutung zugrunde, dass etwas administriert, also verwaltet wird. Homonymie besteht dagegen, wenn kein gemeinsamer Bedeutungsursprung festzustellen ist. Da diese Grenze historisch bedingt allerdings häufig nicht ganz klar zu bestimmen ist, wird oft allgemein vom „Phänomen der Mehrdeutigkeit“ gesprochen (Schwarz und Chur 2007, S. 56). Ein Beispiel dafür ist der Ausdruck *Bank*, der sich zum einen auf eine Sitzgelegenheit und zum anderen auf ein Geldinstitut beziehen kann. Heute haben diese Konzepte keinen Zusammenhang, weshalb der Ausdruck zunächst ein Beispiel für Homonymie darstellt. Jedoch entstammen beide Begriffe dem italienischem *banco* (Tisch)[1] und teilen sich dementsprechend doch den Bedeutungsursprung.

Ausdrücke können sowohl auf lexikalischer als auch auf syntaktischer Ebene ambig sein. Syntaktische oder auch strukturelle Ambiguität tritt immer dann auf, wenn ein Satz auf mehr als eine Weise geparst werden kann (Hirst 1987, S. 131). In diese Kategorie fallen zum Beispiel Sätze, bei denen Konstituenten an unterschiedliche Phrasen im Satz angehängt werden können. Dieses Phänomen tritt beispielsweise bei dem Satz *Ich sehe den Mann mit dem Fernrohr.* auf, bei dem die Phrase *mit dem Fernrohr* entweder an das Subjekt oder das Objekt angehängt werden kann. Neben syntaktischen Mehrdeutigkeiten können auch lexikalische Mehrdeutigkeiten auftreten, wie beispielsweise in dem Satz *Ich sehe das Schloss*. Während die syntaktische Funktion von *Schloss* eindeutig ist, entspringt die Ambiguität in diesem Fall der

[1] https://www.dwds.de/wb/Bank (zuletzt aufgerufen: 24.03.2020)

Mehrfachbelegung verschiedener Bedeutungen auf einer Wortform. Sowohl für syntaktisch als auch für lexikalisch mehrdeutige Ausdrücke gilt, dass beim Lesen (oder Hören) jeweils nur eine Lesart gleichzeitig aktiviert werden kann, sobald die verschiedenen Bedeutungen in antagonistischer Position zueinanderstehen. Das gilt vor allem, wenn der Kontext beide Lesarten zulässt. In diesem Fall wird immer nur eine Lesart gleichzeitig aktiviert (Cruse 2000, S. 31). Schlägt das Parsing mit dieser Lesart jedoch fehl, so wird die andere Lesart ausprobiert.

Sowohl Agirre und Edmonds als auch Neuefeind argumentieren, dass lexikalische Mehrdeutigkeit eine Eigenschaft ist, die prinzipiell alle Wörter besitzen. Diese tritt jedoch häufig nicht zutage, weil die Wörter im Kontext eingebettet sind. Hier wird die intendierte Bedeutung durch umliegende andere sprachliche Ausdrücke definiert. Je mehr Informationen mithilfe des (linguistischen) Kontexts übertragen werden, desto stärker kann die Bedeutung eines Ausdrucks eingegrenzt werden. Fehlen diese Informationen, so ist nicht ersichtlich, welche Bedeutung beabsichtig ist. In diesem Fall entsteht eine Ambiguität im Text (Agirre und Edmonds 2007, S. 8) (Neuefeind 2019, S. 13–14). Erwähnt beispielsweise ein Unternehmen in einer Stellenanzeige, dass „Erfahrung in der Administration“ erwünscht ist, fehlt dem Leser hier Kontextinformation, um zwischen den Konzepten Personaladministration und Serveradministration unterscheiden zu können. In dieser Situation kann weiterer Kontext aus der Anzeige dabei helfen, die Ambiguität zu reduzieren.

Abgesehen von Ambiguität im Text tritt die Mehrdeutigkeit von Begriffen spätestens dann auf, wenn sie isoliert von ihrem Kontext betrachtet werden so wie es bei der Informationsextraktion der Fall ist. Isolation von sprachlichen Ausdrücken tritt auch auf, wenn Wörter in Lexika aufgelistet werden oder bei Suchmaschinen, in die ambige Ausdrücke eingegeben werden (Neuefeind 2019, S. 15). Nach einer vollständigen Loslösung vom Kontext ist es nicht mehr möglich die semantischen Eigenschaften eines Wortes nachzuvollziehen (Cruse 2000, S. 30). Deshalb benötigen wir immer Informationen über den Kontext, um auf die Bedeutung einer sprachlichen Äußerung zu schließen. Dies kann in Datenbanken beispielsweise über Annotationen, die die einzelnen Einträge beschreiben, geschehen.

Nachdem nun exemplarisch dargestellt wurde, dass Kontext für die Bedeutung eines Ausdrucks wichtig ist, wird diskutiert wie Kontext und Bedeutung zusammenhängen können. Eine extreme Position besteht hier in der Meinung, dass Bedeutung ohne Kontext nicht existiert. Demnach würde auch jede Änderung im Kontext eine (zumindest kleine) Änderung der Bedeutung bewirken. Diesen Ansatz nennen wir in dieser Arbeit Bedeutung-pro-Token-Auffassung, weil jede Bedeutung an einen expliziten Wortgebrauch gebunden ist. Cruse bezeichnet diese

Position als unhaltbar (Cruse 2000, S. 30). Außerdem scheint der Ansatz vor allem für eine rechnerbasierte Analyse von Bedeutung nicht praktikabel, da dabei keine aus den bekannten Bedeutungen erzeugten Regeln verwendet werden können, um auf die Bedeutung von neuen Wortgebräuchen zu schließen. Auch wiederspricht ein Ansatz, bei dem jeder Ausdruck seine ganz individuelle Bedeutung hat, der Auffassung über Kategorial- und Partikularkonzepte, wie sie unter 2.1. bereits beschrieben wurden.

Der Theorie von Bedeutung als Kontinuum steht die Theorie der Bedeutungsinventare gegenüber. Ein Bedeutungsinventar ist eine vollständige und unveränderbare Liste von Bedeutungen, die einem Ausdruck zugeordnet sind. Diese Listen finden sich oft in Lexikoneinträgen wieder, bei denen die bekannten Bedeutungen eines Wortes und die jeweiligen Definitionen aufgelistet sind. Auch Thesauri wie beispielsweise *WordNet*, in denen Konzepte miteinander in Relation gesetzt werden, arbeiten mit solchen Bedeutungslisten. Beispielsweise gibt es dort für den Ausdruck *goal* vier verschiedene Bedeutungen: das Ende, die Zielsetzung, das Tor als Gegenstand in einem Sportereignis und das Erzielen eines Punktes[2]. Alle Bedeutungen sind wiederum mit Synsets verlinkt, die wiederum verschiedene Ausdrücke mit der gleichen Bedeutung zusammenfassen. Laut Agirre und Edmonds sollten Bedeutungsinventare so entworfen werden, dass sie klar, konsistent und vollständig sind. Die Vollständigkeit impliziert jedoch, dass Bedeutungsinventare nur für eine begrenzte Domäne entworfen werden können, da man schwerlich immer alle Bedeutungen für einen Ausdruck kennt. Deshalb sprechen Agirre und Edmonds auch davon, dass Bedeutungsinventare bei der Entwicklung auf eine bestimmte Anwendung zugeschnitten werden sollten (Agirre und Edmonds 2007, S. 9). Im Fall der Stellenausschreibungen ist die Domäne klar eingegrenzt. Alle Terme im Korpus treten nur im Arbeitsmarkt-Kontext auf und beziehen aus diesem Rahmen ihre Bedeutung. Aus diesem Grund ist es allerdings auch schwierig, Taxonomien wie beispielsweise *WordNet* oder das deutschsprachliche Pendant *GermaNet*[3] zu verwenden, die nicht im Kontext von Stellenanzeigen entwickelt wurden. Allerdings gibt es spezielle Kompetenz-Taxonomien, die unter 4.2. vorgestellt werden.

Zwischen diesen beiden Extremen findet eine abgeschwächte Variante der Bedeutung-pro-Token-Auffassung am meisten Zustimmung. Denn bildet man die einzelnen Wortgebräuche (*word usages*) auf Grundlage ihres jeweiligen Kontexts in einem Kontinuum ab, so entstehen Regionen von unterschiedlich starker Dichte. Regionen mit hoher Dichte können laut dieser

[2] http://wordnetweb.princeton.edu/perl/webwn?s=goal (zuletzt aufgerufen: 24.03.20)
[3] http://www.sfs.uni-tuebingen.de/GermaNet/ (zuletzt aufgerufen: 24.03.20)

Auffassung zu einer Bedeutung zusammengefasst werden (Agirre und Edmonds 2007, S. 9). Das bedeutet gleichzeitig, dass die Bedeutung eines Ausdrucks auch unter (kleinen) Veränderungen im Kontext stabil bliebt (Cruse 2000, S. 30). Wie Bedeutung durch Informationen aus dem Kontext in einem Kontinuum abgebildet werden kann, wird in Kapitel 2.4. näher erläutert. Sobald man Bedeutungen als Cluster in einem kontinuierlichen Raum auffasst, ergeben sich weitere Möglichkeiten für den Umgang mit neuen Wortgebräuchen. Zum Beispiel kann man die Verknüpfung zwischen einem Ausdruck im Text und einer bekannten Bedeutung auch als Klassifikationsaufgabe sehen. Denn dabei geschieht nichts anderes als, dass man den neuen Ausdruck einem bestehenden Cluster zuordnet und dem Wort somit die Bedeutung, die das Cluster repräsentiert, zuweist. Wie eine solche Klassifikation durchgeführt werden kann, wird in Kapitel 5 erläutert. Neben der Mehrdeutigkeit als Relation zwischen Ausdruck und Inhalt kann auch das Phänomen der Synonymie auftreten, welches im Folgenden vorgestellt wird.

2.3. Synonymie

Genau wie eine Wortform mit verschiedenen Konzepten verbunden sein kann, lassen sich einem Konzept manchmal verschiedene Wortformen zuordnen. Bei diesem Phänomen, das als Synonymie bezeichnet wird, herrscht eine „Bedeutungsgleichheit zwischen Wörtern“ (Schwarz und Chur 2007, S. 54). Ob zwei Wörter synonym sind lässt sich beispielsweise erkennen, indem man prüft, ob die Bedeutung eines Satzes nach dem Ersetzen eines dieser Wörter noch deckungsgleich ist. In Bezug auf die Merkmalshypothese sollten außerdem die Merkmalsmengen der beiden Bedeutungen identisch sein. Vollständig synonyme Ausdrücke gibt es selten, oft ist zumindest die Konnotation der Wörter unterschiedlich, wie zum Beispiel bei *Pferd* und *Gaul*. Neben den denotativen Merkmalen, die die „semantische Grundbedeutung eines Wortes angeben“, übermitteln konnotative Merkmale „zusätzliche, meist pejorative, emotional gefärbte Informationen“ (Schwarz und Chur 2007, S. 54–55). Diese kommen besonders bei Wertungen zum Tragen und sind zum Beispiel für die Sentiment Analysis relevant. Konnotative Merkmale sind individuell, durch Sprachgemeinschaften oder von bestimmten Stilebenen geprägt. Für die Kategorisierung von Kompetenzen in Stellenanzeigen sind zunächst vor allem die denotativen Merkmale der Wörter interessant.

Neben den typischen Beispielen für Synonymie, die in der Literatur aufgeführt werden, sind für das Concept Mining mit Stellenausschreibungen auch orthografische Varianten von Kompetenzen zu beachten. Weil für die Verfasser von Stellenanzeigen kein standardisiertes Wörterbuch vorliegt, können Kompetenzausdrücke unterschiedlich geschrieben werden. Dies

betrifft zum einen Ausdrücke, die entweder als Mehrwortausdrücke oder mit Bindestrich geschrieben werden (*MS-Office* vs. *MS Office*) und zum anderen unterschiedliche von der Rechtschreibung zugelassene Varianten (*graphisches Zeichnen* vs. *grafisches Zeichnen*). Einige Fälle werden hier bereits durch die Taxonomien abgebildet (siehe 4.2.), allerdings benötigt das Framework zusätzlich Verfahren, die noch neu aufkommende Orthografien mit aufnehmen (siehe 4.1.1.).

2.4. Vektorraummodelle und die distributionelle Hypothese

Eine Möglichkeit, um nun die Beziehungen zwischen verschiedenen Konzepten oder verschiedenen Ausdrücken zu untersuchen, ist die zu untersuchenden Elemente in einem Raum anzuordnen. In einem sogenannten Vektorraummodell können beispielsweise Wörter als Punkte im mehrdimensionalen Raum verteilt werden. Punkte, die nah beieinander liegen, deuten semantische Nähe an, während weit entfernte Punkte semantische Distanz bedeuten (Turney und Pantel 2010, S. 141). Sahlgren spricht auch von einem *word-space model* und meint damit ein rechnerbasiertes Modell, das Bedeutung modelliert (Sahlgren 2006, S. 9). Sein Anspruch an das Modell ist, dass es vollkommen datengetrieben entsteht und dadurch kaum Vorannahmen über Sprache in die Modellierung von Bedeutung mit einfließen. Demnach müsste sich das Modell anpassen, sobald sich die Bedeutung von Ausdrücken über die Zeit verändert. Außerdem sehe das Modell anders aus, wenn man eine andere Datenquelle, zum Beispiel Daten aus einer anderen Domäne, verwenden würde (Sahlgren 2006, S. 11). Das würde bedeuten, dass es für die semantische Analyse von Stellenanzeigen sinnvoll wäre, Modelle zu verwenden, die innerhalb der Stellenausschreibungen-Domäne entwickelt wurden. Die Annahme wäre nämlich, dass Modelle, die auf einer anderen Grundlage erstellt wurden, nicht die für Stellenanzeigen spezifischen Bedeutungen abbilden. Um diese Annahme zu überprüfen, werden unter 4.3. verschiedene Vektorraummodelle gegenübergestellt.

Sowohl das *word-space model* als auch andere Vektorraummodelle werden oft durch eine Matrix dargestellt, die alle Vektoren des jeweiligen Modells abbildet. In Modellen, die Wörter als Untersuchungsgegenstände haben, wird jedes Wort durch einen Vektor repräsentiert. Um hier genauer auf Wortvektoren eingehen zu können, müssen Wörter allerdings zuerst noch definiert werden. Sahlgren referenziert mit „Wort“ auf eine durch Leerzeichen abgegrenzte Zeichenkette, welche auf die Grundform normalisiert ist (Sahlgren 2006, S. 12). Demnach verwendet er in seinem Vektorraummodell lemmatisierte Types als Basis im Gegensatz zu anderen Modellen, die auf Grundlage von Tokens aufgebaut werden (Turney und Pantel 2010, S. 151). Sahlgren spricht auch an, dass seine Definition sich auf Textdaten bezieht und

beispielsweise für die Untersuchung von Sprachdaten angepasst werden müsse. Auch gebe es oft Fälle, in denen durch Leerzeichen abgetrennte Zeichenketten nicht unbedingt die klügste Einheit als Grundlage für ein semantisches Modell sind. Beispielsweise sollte das *word-space model* entsprechend angepasst werden, um einzelne Morpheme oder Mehrwortausdrücke zu untersuchen. Dies sei allerdings nur eine Entscheidung während der Vorverarbeitung, die die wesentliche Funktionsweise des Modells nicht beeinflusst (Sahlgren 2006, S. 12).

Vektorraummodelle sollen die Bedeutungen von Wörtern in einem geometrischen Raum darstellen. Diese geometrische Metapher für Bedeutung definiert Sahlgren wie folgt:

> Meanings are locations in a semantic space, and semantic similarity is proximity between the locations. (Sahlgren 2006, S. 19)

Die Metapher ist laut Sahlgren keineswegs zufällig oder arbiträr. Vielmehr kommt sie daher, dass wir als „verkörperte Wesen" abstrakte Konzepte wie semantische Ähnlichkeit durch Mittel aus unserer räumlich-zeitlichen Erfahrung wie zum Beispiel räumliche Distanz veranschaulichen wollen (Sahlgren 2006, S. 18). Dabei gilt die Metapher von Nähe für Ähnlichkeit (*similarity-as-proximity metaphor*) nicht nur in der Semantik (Sahlgren 2006, S. 19). Die Punkte, die beispielsweise Wörter repräsentieren, werden in einem n-dimensionalen Vektorraum positioniert, wobei die Anzahl der Dimensionen durch das jeweilige Modell festgelegt wird (Sahlgren 2006, S. 18). Dabei kann jede Dimension Einfluss auf die Ähnlichkeit bzw. die Unähnlichkeit von zwei Wörtern nehmen. Eine Erläuterung, wie die Position der einzelnen Punkte berechnet wird, folgt in einem anschließenden Beispiel.

Beispielsweise könnte man die Wörter *uncool*, *ätzend*, *super* und *grandios* auf einer Achse auftragen je nachdem, wie häufig man die Wörter täglich ausspricht. Je häufiger ein Wort am Tag beim Sprechen verwendet wird desto höher wird es auf der Achse aufgetragen. Das Ergebnis könnte dann so aussehen:

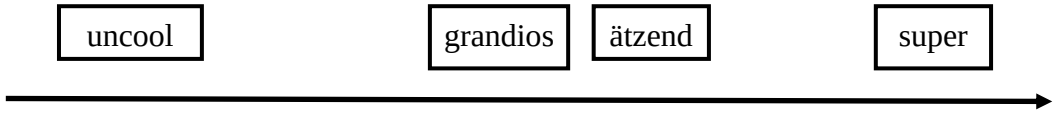

Abbildung 2: eindimensionales Vektorraummodell

In dieser Darstellung befinden sich *grandios* und *ätzend* in räumlicher Nähe während *grandios* und *super* weiter voneinander entfernt sind. Nimmt man jedoch eine weitere Dimension hinzu,

die ausdrückt, ob die Wörter etwas positives oder etwas negatives ausdrücken, so ändert sich die Anordnung beispielsweise so:

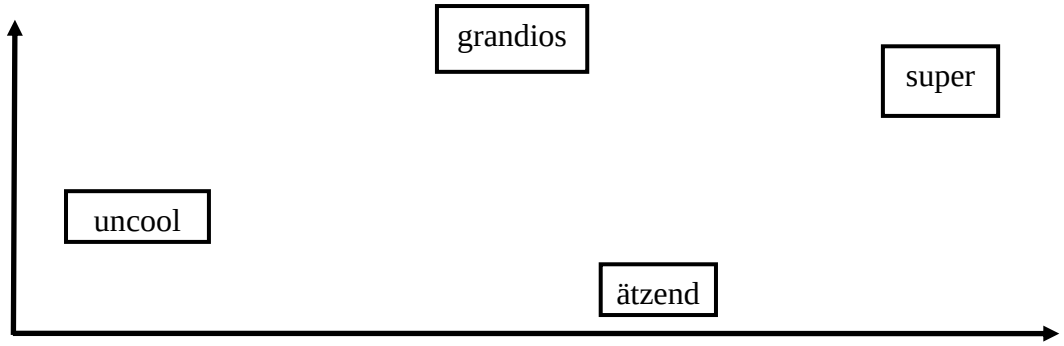

Abbildung 3: zweidimensionales Vektorraummodell

Die Nähe zwischen *grandios* und *ätzend* wirkt nun nicht mehr so groß, da die beiden Begriffe sehr unterschiedliche Haltungen ausdrücken. Wenn man die zweite Achse isoliert betrachtet, gruppieren sich nun eher *uncool* und *ätzend* sowie *super* und *grandios*. Nähme man nun eine weitere Dimension hinzu, die beispielsweise abbildet, die häufig man die Begriffe täglich schreibt, könnte sich das Bild der räumlichen Distanzen wieder verschieben. Jede Dimension kann also die Nähe bzw. Distanz von Punkten mitunter erheblich beeinflussen. Sowohl in diesem Beispiel als auch allgemein in Bezug auf Vektorraummodelle ist wichtig zu erwähnen, dass nicht die absolute Verortung eines Punkts im Raum entscheidend ist. Die Leistungsfähigkeit des Ansatzes kommt erst zum Tragen, wenn man die geometrische Relation zwischen verschiedenen Punkten untersucht (Sahlgren 2006, S. 19). So spielt die absolute Verortung der Wörter nur im Rahmen der Operationalisierung eine Rolle. Im Beispiel könnte man auch die Achsen vertauschen oder die Skalen invertieren, entscheidend ist nur das Muster der räumlichen Distanzen, das dabei entsteht.

2.4.1. Vektoren im Vektorraum erzeugen

Noch ist nicht geklärt wie die datengetriebenen Vektoren ermittelt werden sollen. Im vorangegangenen Beispiel dienten dazu Frequenzanalysen (Häufigkeit pro Tag) sowie subjektive Einschätzungen (positive und negative Konnotationen). Im Bereich der Korpuslinguistik gib es allerdings noch weitere Methoden zur Berechnung der Wortvektoren, die im Folgenden vorgestellt werden. Dazu ist es zunächst wichtig zu verstehen was diese Vektoren genau repräsentieren sollen. Dies wird zunächst am Beispiel von Wortvektoren

gezeigt. Die Erzeugung von Dokumentenvektoren wird anschließend erläutert. Die Bedeutung eines Ausdrucks ergibt sich wie bereits erwähnt aus seinem Kontext (siehe 2.1.). Bei der Erzeugung von Wortvektoren spielt vor allem der linguistische Kontext eine Rolle. Dies wird auf der von Zellig Harris entwickelten distributionellen Methodik begründet, bei der davon ausgegangen wird, dass sich Wörter von gleichen linguistischen Klassen, wie beispielsweise Morpheme, Phoneme oder Wortarten, innerhalb einer Sprache (*parole*) ähnlich verteilen. Harris wollte dieses Phänomen nutzen, und Wörter entsprechend ihres Verhaltens gruppieren (Sahlgren 2006, S. 22). Daraus resultiert schließlich die distributionelle Hypothese, die vermutet, dass Worte mit ähnlichen Verteilungseigenschaften auch ähnliche Bedeutungen haben müssen (Sahlgren 2006, S. 21). Wenn beispielsweise die Wörter *blaue* und *rote* immer vor dem Wort *Blume* stehen, haben sie eine ähnliche Verteilung in Bezug auf ihren Kontext und demnach eine ähnliche Bedeutung. Turney und Pantel sprechen bei dem Zusammenhang von *blaue* und *rote* auch von einer paradigmatischen Beziehung, die häufig auftritt, wenn eine taxonomische Ähnlichkeit besteht. In diesem Fall gehören beide der Gruppe der Farben an. Wörter, die in einer paradigmatischen Beziehung stehen, treten selten zusammen in einem Satz auf, weil sie oft die gleiche syntaktische Position einnehmen. Die Beziehung zwischen *blaue* und *Blume* hingegen beschreiben Turney und Pantel als syntagmatisch, weil diese Wörter häufig als Nachbarn zusammen in einem Satz auftauchen. Diese Art von Beziehung deute oft auf eine semantische Verbundenheit hin (Turney und Pantel 2010, S. 150). Um paradigmatische Beziehungen anhand von Vektoren abbilden zu können, bietet es sich an, Wörter nur aufgrund ihrer Kontextwörter darzustellen. Dafür kann man beispielsweise die Wörter im Kontext zählen. Sollten zwei sprachliche Ausdrücke bei jedem Wortgebrauch ähnliche Nachbarn im Text haben, so werden ihre Vektoren in den einzelnen Dimensionen ähnlich ausgeprägt sein und dementsprechend näher beieinander liegen als Vektoren von Ausdrücken, bei denen die Nachbarn nicht oder weniger übereinstimmen. Die Vektoren von *rote* und *blaue* sollten also idealerweise ähnlicher sein als die Vektoren von *rote* und *Blume.* Die Auswahl der Nachbarn im Text kann dabei variieren. Je nach Ansatz werden hier nur das direkt vorhergehende und das direkt nachfolgende Wort als Kontext gewählt, in anderen Fällen wir das Fenster breiter gesetzt (Turney und Pantel 2010, S. 170).

Oft werden Vektorraummodelle auch verwendet, um die semantische Beziehung zwischen Dokumenten zu untersuchen, zum Beispiel in einer Term-Dokument-Matrix. Hier geht es nicht mehr um die Beziehung zwischen einem Wort und den umliegenden Wörtern, sondern um die An- bzw. Abwesenheit von Wörtern in Dokumenten. Der Begriff Dokument steht hier lediglich für eine Einheit, die mehrere Wörter zusammenfasst. Damit können einzelne Sätze, Tweets,

Briefe oder auch ganze Romane gemeint sein. Dokumenten mit ähnlichen Vektoren sagt man auch hier semantische Ähnlichkeit nach, da die Wörter in ihnen ähnlich verteilt sind. Dabei wird die Reihenfolge der Wörter allerdings nicht berücksichtigt. Gleichzeitig entstehen in einer Term-Dokument-Matrix auch wieder Wortvektoren, und zwar weil für jedes Wort die Information vorliegt, in welchen Dokumenten es vorkommt. Dadurch lassen sich auch Aussagen über die Verteilung der einzelnen Wörter treffen und die Verteilungen von zwei Wörtern über verschiedene Dokumente hinweg vergleichen (Turney und Pantel 2010, S. 142).

Sowohl mit Blick auf die Wort-Kontext-Matrix als auch die Wort-Dokument-Matrix ist wichtig festzuhalten, dass die Definition eines Wortes entscheidend ist. Je nach Zielsetzung kann es beispielsweise sinnvoll sein, Tokens in ihrer flektierten Form zu einem Vektor bzw. einer Dimension zusammenzufassen. Oft werden diese aber zuvor normalisiert, zum Beispiel indem sie auf ihre Grundform zurückgeführt werden. Sowohl bei flektierten Wortformen als auch bei lemmatisierten Wortvektoren wird nicht auf das Problem eingegangen, das mehrere Lexeme durch einen sprachlichen Ausdruck zusammengefasst werden. Dieser Umstand führt zwangsläufig dazu, dass diese Vektoren nicht ohne weiteres dazu genutzt werden können, um die einzelnen Lexeme eines Ausdrucks voneinander abzugrenzen. Eine Lösung dieses Problems könnte sein, die Bedeutungen der Ausdrücke jeweils zu annotieren, sodass für jede Bedeutung ein eigener Vektor gebildet werden kann. Dies ist allerdings kein leichtes Unterfangen, da einerseits manuell Korpora annotiert werden müssen und andererseits die Bedeutungen vor allem in Bezug auf die Variabilität von Sprache nicht immer eindeutig abzugrenzen sind (siehe 2.2.).

2.4.2. Word Embeddings

Neben den beschriebenen Vektorraummodellen finden vor allem Word Embeddings in den letzten Jahren in der Computerlinguistik Anklang. Sie stellen gewissermaßen eine Subklasse der Vektorraummodelle dar und beruhen ebenfalls ausschließlich auf Kontextinformationen (Li et al. 2018, S. 270). Die Wortvektoren werden dabei mithilfe von neuronalen Netzen erzeugt. Eine der bekanntesten Embedding-Architekturen ist Word2Vec, bei dem es darum geht, ein Wort auf einem Vektor abzubilden, sodass semantische Beziehungen zwischen Wörtern auf geometrische Operationen übertragen werden können. Word2Vec nutzt für die Erstellung des Vektorraummodells einen sogenannten *‘fake‘ task*. Statt ein neuronales Netz darauf auszurichten, dass es direkt Wortvektoren berechnet, wird zum Beispiel die Aufgabe gestellt, von einem Lückentext auf das jeweils fehlende Wort zu schließen. Diese Aufgabe wird auch als *Continous Bag-of-Words* (CBOW) bezeichnet (siehe Abbildung 4). Mikolov et al. schlagen

als Eingabedaten für das neuronale Netz jeweils vier Wörter vor und vier Wörter nach dem gesuchten Wort vor (Mikolov et al. 2013a). Als Ausgabe liefert das Neuronale Netz dann das Wort, das für die freigelassene Position am wahrscheinlichsten ist.

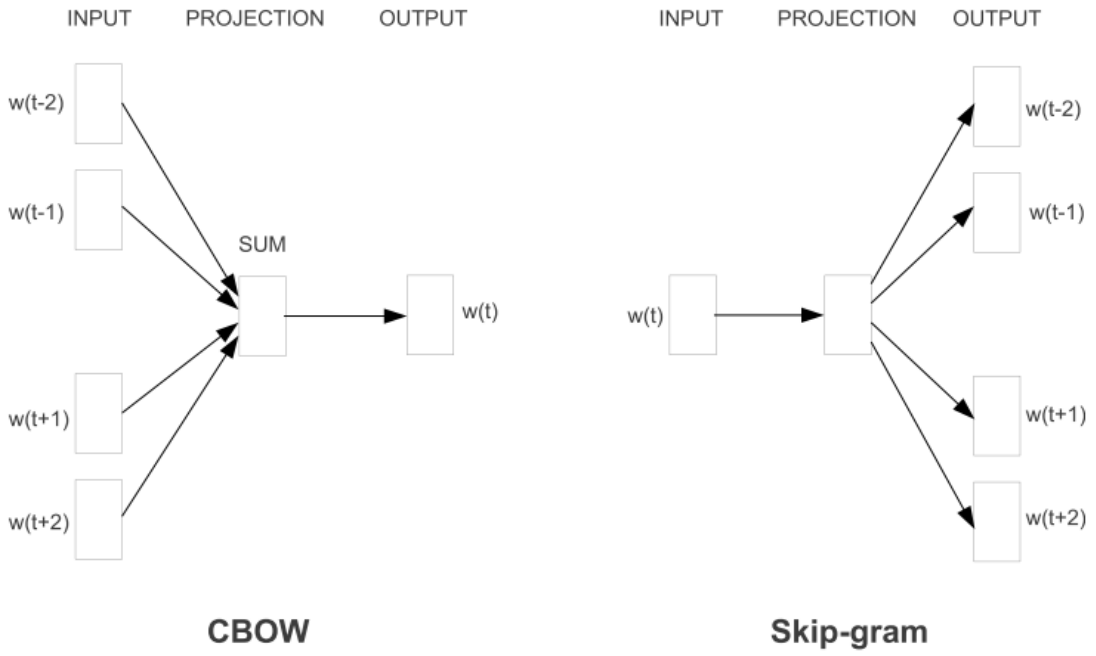

Abbildung 4: Architektur von CBOW und Skip-gram (Mikolov et al. 2013b)

Eine andere Word2Vec-Architektur, die *Skip-gram-Model* genannt wird, sieht es vor, für ein bekanntes Wort die Nachbarn vorauszusagen. Um weiter entferntere Wörter weniger stark zu gewichten, werden diese auch seltener in die Trainingsdaten aufgenommen (Mikolov et al. 2013a). Sowohl Skip-gram- also auch CBOW-Modelle werden mithilfe von Wort-Kontext-Paaren aus Korpora trainiert bis die Voraussagen möglichst akkurat sind. In beiden Fällen hat das Neuronale Netz ein Hidden Layer, dessen Größe als Hyperparameter festgelegt wird. Nach Beendigung des Trainings enthält im Fall von Skip-Gram die Matrix von genau diesem Hidden Layer die eigentlichen Wortvektoren. Jeder Vektor ist dabei so lang wie die Größe des Hidden Layers. Die Vektoren können nach dem Training genutzt werden, um semantische Beziehungen zwischen den einzelnen Wörtern durch geometrische Operationen abzubilden. Diese Funktion wird in Kapitel 4.3. genutzt, um die semantischen Beziehungen von Wörtern in Stellenanzeigen mit den Beziehungen von Wörtern in Wikipedia-Artikeln zu vergleichen. Genau wie andere Vektorraummodelle ist Word2Vec nicht in der Lage die Lexeme in einem Ausdruck zu differenzieren. Es gibt aber innovative Methoden, mit denen Word Embeddings erweitert werden können, um mit den Vektoren Bedeutungen und keine Wörter abzubilden, die im folgenden Kapitel präsentiert werden.

3. Forschungsstand

3.1. Methoden zur *Word Sense Disambiguation*

Word Sense Disambiguation (WSD) ist in der Computerlinguistik ein viel diskutiertes Thema. Iacobacci et al. bezeichnen Disambiguierung sogar als „eine der ältesten Aufgaben im Bereich von Natural Language Processing und Künstlicher Intelligenz“ (Iacobacci et al. 2016, S. 897). Dementsprechend breit ist das Spektrum der Ansätze zur WSD. Auf der einen Seite stehen dabei Klassifikationswerkzeuge, die überwacht arbeiten, wie beispielsweise das von Zhong und Ng entwickelte *It Makes Sense* (IMS). Sie wollen dabei für alle Inhaltswörter, also Nomen, Verben, Adjektive und Adverbien, die Bedeutung des jeweiligen Wortes anhand von (linguistischen) Merkmalen des Kontexts klassifizieren. Die in Frage kommenden Bedeutungen entnehmen sie dabei *WordNet*, einem englischsprachigen Thesaurus (Zhong und Ng 2010, S. 79). Für die Klassifikation eines Tokens ziehen sie dabei die Wortarten der drei vorhergehenden und nachfolgenden Wörter, die lemmatisierten umliegenden Wörter selbst und lokale Kollokationen heran (Zhong und Ng 2010, S. 79). Diese Merkmale werden genutzt, um das ambige Wort mithilfe von Support Vector Machines (SVM) einer der zur Auswahl stehenden Bedeutungen zuzuordnen (Zhong und Ng 2010, S. 78). Die Klassifikation führen Zhong und Ng auf den Trainingsdaten, die für die SensEval- und SemEval-Wettbewerbe genutzt werden, durch. Sie erreichten damit eine Accuracy von 65.3% für SensEval-2 bzw. 72.6% für SensEval-3 und sind damit etwas besser als die jeweiligen Gewinnersysteme der Wettbewerbe (Zhong und Ng 2010, S. 81).

Agirre und Martínez verfolgen eine andere Strategie. Sie kritisieren überwache Ansätze vor allem dafür, dass sie auf Bedeutungsinventaren basieren. Diese seien zu starr, weil sie Bedeutungen in diskrete Kategorien einteilen, was der Ansicht von vielen Lexikographen und Semantikern wiederspricht. Stattdessen verfolgen sie den Ansatz, dass sich Bedeutung über Cluster in einem kontinuierlichen Raum abbilden lässt, wie es auch bereits im vorherigen Kapitel (2.2) vorgestellt wurde (Agirre et al. 2006, S. 585). Sie wollen diese Cluster identifizieren, indem sie in einem Kookkurrenz-Graphen zentrale Knoten, die sogenannten Hubs identifizieren, welche jeweils die Bedeutung eines Worts repräsentieren (Agirre et al. 2006, S. 587). Durch dieses Verfahren ist es möglich die Bedeutungen aus den Texten heraus zu induzieren. Dies hat vor allem den Vorteil, dass Bedeutungen nicht von außen aufgelegt werden, ähnlich wie in Zusammenhang mit dem *word-space model* beschrieben (siehe 2.4.). Die Abbildung von Bedeutungen in Clustern entspricht der Auffassung von Bedeutung, die sich im Kontinuum in Regionen von unterschiedlicher Dichte sammelt (siehe 2.2).

Vor allem in den letzten zehn Jahren hat die Verwendung von Embeddings (siehe 2.4.2.) in die WSD-Forschung Einzug gehalten. Beispielsweise verwenden Rothe und Schütze bestehende Word-Embeddings in Kombination mit Thesauri, wie sie auch Zhong und Ng verwenden. Jedoch nutzen Rothe und Schütze nicht nur die Einträge in beispielsweise *WordNet* als feste Knoten für die Evaluation ihres Systems (Rothe und Schütze 2015, S. 1793). Vielmehr geht es in ihrem Projekt *AutoExtend* darum, für alle Lexeme im Thesaurus einzelne Vektoren zu berechnen und auf Grundlage dieser wiederum Vektoren für Synsets zu erstellen. Ihr Ziel ist es die trainierten Word-Embeddings so zu erweitern, dass im Vektorraum der Wörter alle Vektoren für Lexeme und Konzepte gleichzeitig abgebildet werden können, um so geometrische Operationen zwischen Lexemen, Konzepten und Ausdrücken zu ermöglichen (Rothe und Schütze 2015, S. 1794).

Diesen neueren von hoher Rechenleistung unterstützen Ansätzen stehen wiederum Verfahren gegenüber, die sich semantische Eigenschaften von Texten zum Vorteil machen. Beispielsweise verwendete Resnik Selektionspräferenzen von Verben zur Disambiguierung. Der Begriff der Selektionspräferenz oder Selektionsbeschränkung entspringt der Valenzgrammatik und beschreibt dort zunächst lediglich die Tatsache, dass Wörter auf semantischer Ebene nur unter bestimmten Bedingungen miteinander kombiniert werden können (Metzler Lexikon Sprache, 5. Auflage, s.v. „Selektionsbeschränkung“). Die Valenzgrammatik beruht auf Tesnières Prinzip der Dependenz und steht somit im Gegensatz zur Konstituentengrammatik. Wörter werden hier nicht in Phrasen zusammengefasst, sondern inhaltsbezogen verknüpft. Im Zentrum eines Satzes steht dabei das Verb, welches auch als Valenzträger bezeichnet wird, das eine bestimmte Anzahl an anderen Wörtern benötigt, damit der Satz vollständig ist. Diese anderen Wörter werden auch Satelliten genannt (Metzler Lexikon Sprache, 5. Auflage, s.v. „Valenzgrammatik“). Verben können so in verschiedene Kategorien unterteilt werden, je nachdem welche anderen Wörter sie benötigen. Die Valenzgrammatik wurde mit der Zeit um einen semantischen Aspekt erweitert, der sich mit den zuvor genannten Selektionsbeschränkungen befasst. Damit ist gemeint, dass neben der Erfüllung der syntaktischen Vorgaben auch eine inhaltliche Harmonie zwischen Valenzträger und Satellit herrschen muss. Zum Beispiel ist der Satz *Ich schreibe dir ein Auto.* zwar syntaktisch korrekt, jedoch hat er aus semantischer Sicht keinen Wahrheitswert. Das liegt daran, dass das Verb *schreiben* als thematische Rolle nicht jedes beliebige Wort zulässt. Nur Dinge, die geschrieben werden, wie beispielsweise ein Brief, ein Lied, eine Nachricht etc. können die Stelle ausfüllen, sodass der Satz semantisch korrekt wird. Die Valenztheorie wurde zunächst nur für Verben

formuliert, jedoch lässt sich das Prinzip auch auf andere Wortarten, wie beispielsweise Nomen, ausweiten.

Im Rahmen der Disambiguierung geht es nun nicht darum, ob ein bestimmter Satz semantisch korrekt ist. Vielmehr soll der Zusammenhang zwischen Selektionsbeschränkungen und Wortbedeutungen, der bereits vielfach untersucht wurde, zur Disambiguierung genutzt werden. Die Hypothese ist hierbei, dass sich die taxonomiellen Kategorien der Satelliten mit der Bedeutung eines Valenzträgers ändert. Resniks Ansatz unterscheidet sich von anderen dadurch, dass keine bedeutungsannotierten Korpora im Training notwendig sind. Diese Entscheidung begründet er unter anderem damit, dass Bedeutungsannotationen häufig fehlerhaft sind, weil selbst menschliche Annotatoren oft keine einheitliche Annotation über ein Korpus hinweg erstellen können (Resnik 1997, S. 52). Resnik wählt für die Disambiguierung einen probabilistischen Ansatz, bei dem die Kookkurrenz zwischen einem Valenzträger und einer Konzeptklasse in einer Taxonomie untersucht wird (Resnik 1997, S. 52). Beispielsweise untersucht er für das Verb *buzz* (deutsch: *summen, sirren*), aus welcher Kategorie die Subjekte, die damit verbunden sind, stammen. Daraus ergibt sich, dass die Wahrscheinlichkeit für ein Insekt in dieser Position wesentlich höher ist als die für eine Person. Die Klassen der Satelliten werden dabei wie bei *AutoExtend* oder *It Makes Sense* aus *WordNet* entnommen. Das Prinzip der Selektionsbeschränkung könnte auch im Kontext der Stellenanzeigen hilfreich sein. Beispielsweise könnte so der Bedeutungsunterschied zwischen *Beratung von Personal* und *Bedeutung von Patienten* abgebildet werden, weil *Personal* und *Patienten* unterschiedlichen semantischen Konzepten zugeordnet sind.

3.2. Natural Language Processing in der Arbeitsmarktforschung

Disambiguierung spielt in der Auswertung von Stellenausschreibungen bisher nur eine untergeordnete Rolle. Jedoch gibt es viele Forschungsprojekte, die NLP-Verfahren allgemein auf Stellenanzeigen anwenden. Vor allem im Bereich der Computational Social Sciences finden Bestrebungen statt, die Erkenntnisse aus der Computerlinguistik zu nutzen, um unstrukturierte Stellenanzeigen für weitere Analysen aufzubereiten. Zum Beispiel haben Calanca et al. in ihrer Studie Soft Skills in Stellenanzeigen näher untersucht (Calanca et al. 2019). Soft Skills haben in der Arbeitsmarktforschung eine gesonderte Rolle, weil sie bereichsübergreifend eine Rolle spielen. Gleichzeitig sind sie allerdings nicht so leicht zu quantifizieren wie die sogenannten Hard Skills, also fachliche Kompetenzen wie technische Fähigkeiten oder berufliche Qualifikationen (Calanca et al. 2019, S. 1). Innerhalb der Soft Skills-Forschung gibt es beispielsweise Projekte, in denen untersucht wird, wie bestimmte Formulierungen von Soft

Skills Gender-Stereotypen unterstützen. Um in diesem Bereich quantitative Analysen durchzuführen, ist es notwendig Soft Skills strukturiert aus Stellenanzeigen zu extrahieren (Calanca et al. 2019, S. 2). Jedoch stießen Calanca et al. bei einer solchen Extraktion auf das Problem, dass Adjektive wie *flexibel* entweder auf den/die Bewerber*in oder auf das Unternehmen bezogen sein können. Im ersten Fall würde der Ausdruck einen Soft Skill repräsentieren, der für die quantitative Untersuchung aufgenommen werden soll. Im zweiten Fall wäre der String allerdings für die Analyse von Soft Skills irrelevant, da er sich nicht auf den/die Bewerber*in bezieht. Alle Ausdrücke, die möglicherweise auf einen Soft Skill referenzieren, müssen also dahingehend klassifiziert werden, ob sie für weitere quantitative Analysen extrahiert werden sollen oder nicht (Calanca et al. 2019, S. 6). Bei der hier vorliegenden Arbeit tut sich ein ähnliches Problem auf, das sich nicht nur auf Soft Skills beschränkt. Ausdrücke, die als potenzielle Kompetenzen extrahiert wurden, können sich hier ebenfalls auf Unternehmen beziehen. Außerdem kann es sich um Kompetenzen handeln, die erst im Beruf erlernt werden sollen und nicht von vornherein mitgebracht werden müssen. Also können falsch positive Extraktionen nicht nur durch die unter 2.2. vorgestellten Ambiguitäten entstehen. Diese Fehler gilt es ebenfalls durch weitere Maßnahmen zu minimieren.

Ein anderes Klassifikationsproblem behandelt Stohr in seiner Dissertation zur Digitalisierung auf dem Arbeitsmarkt (Stohr 2019). Er erprobt hier verschiedene Ansätze, um Stellenanzeigen in Bezug darauf zu klassifizieren, wie stark die darin enthaltenen Tätigkeiten durch automatisierte Verfahren ersetzt werden können. Die Herausforderung besteht hier also darin, für eine bestimmte Stellenausschreibung vorherzusagen, welches Substituierbarkeitspotenzial vorliegt. Dieses Potenzial gibt an, wie viele Tätigkeiten in einem Beruf durch digitale Technologien schon jetzt übernommen werden können (Dengler und Matthes 2015). Für die Klassifikation wurden verschiedene Algorithmen aus dem Bereich des Machine Learnings verwendet. Die Klassifikation findet dabei anhand von signifikanten Schlüsselwörtern statt, die pro Stellenanzeige durch den Einsatz von Assoziationsmaßen, wie beispielsweise *Pointwise Mutual Information*, identifiziert wurden (Stohr 2019, S. 111). Die Stellenanzeigen werden demnach durch ihre Schlüsselwörter repräsentiert und anhand dieser in die Substituierbarkeits-Klassen kategorisiert. Allerdings arbeitet dieser Ansatz ausschließlich auf der Ausdrucksebene und behandelt dementsprechend keine potenziellen Ambiguitäten.

4. Computerlinguistische Verarbeitung von Stellenausschreibungen

4.1. Das Projekt „Qualifikationsentwicklungsforschung“

Neben den soeben beschriebenen Projekten im Bereich der Arbeitsmarktforschung, ist diese Arbeit ebenfalls als Teil eines Projekts entstanden, das sich mit der Extraktion von relevanten Termen aus Stellenanzeigen befasst. Die Extraktion semantischer Konzepte kann Anwendungen in vielen Bereichen verbessern, indem sie überall, wo Informationsextraktion betrieben wird, den Informationsgehalt der extrahierten Terme erhöht. Dies gilt auch bei der semantischen Aufbereitung von Stellenanzeigen. Stellenausschreibungen bieten einen guten Proxy für die Abbildung des Arbeitsmarkts, spiegeln sie doch wider, welche Ressourcen unter welchen Bedingungen in welchen Bereichen auf dem Arbeitsmarkt gesucht werden. In dem Projekt *Qualifikationsentwicklungsforschung* wird seit 2015 ein Framework[4] entwickelt, das dazu dient, relevante Informationen aus Stellenanzeigen zu extrahieren und zu kategorisieren. An der Entwicklung sind sowohl das Bundesinstitut für Berufsbildung (BIBB) als auch das Institut für Digital Humanities der Universität zu Köln (IDH) beteiligt. Ziel der Kooperation ist es ein Framework erarbeitet, das eine strukturiertere Aufbereitung von Stellenanzeigen ermöglicht, indem es relevante Informationen in den Ausschreibungen entdeckt und extrahiert. Als relevante Informationen sind hier die im Beruf auszuführenden Tätigkeiten und die zu verwendenden Arbeitsmittel sowie die vom Arbeitnehmer zu erfüllenden Kompetenzen definiert. Die Extraktion dieser Terme aus einem Korpus von Stellenanzeigen bietet die Möglichkeit, die Anforderungen auf dem Arbeitsmarkt zu einem bestimmten Zeitpunkt zu analysieren. Somit können Anwendungen für ein Arbeitsmarkt-Monitoring entworfen werden, die nahezu zeitgleich mit dem Erscheinen von neuen Anzeigen aktuelle Trends beobachten können. Solche Monitoring-Aufgaben wurden bisher durch Erhebungen auf dem Arbeitsmarkt übernommen, bei denen beispielsweise Unternehmen gezielt nach ihren Anforderungen befragt wurden. Dieses Verfahren hat allerdings zwei Nachteile: Zum einen ist die quantitative Erhebung in Betrieben vergleichsweise teuer und die Rücklaufquoten von Ergebnissen gering. Aus diesen beiden Gründen resultiert gleichzeitig der zweite Nachteil: Die Umfragen liefern keine repräsentativen Ergebnisse. Diese Probleme kann man mit der Auswertung von Online-Stellenanzeigen verringern, da die Anzeigen flächendeckend(er) verfügbar sind und keine zusätzliche Erhebung parallel zum Geschehen auf dem Arbeitsmarkt notwendig ist. Das in der

[4] https://github.com/spinfo/quenfo

Kooperation entwickelte Framework wird im Folgenden kurz vorgestellt. Die in dieser Arbeit entwickelte Disambiguierung soll das Framework als zusätzliches Modul ergänzen.

4.1.1. Das Framework

Die Arbeit im Projekt hat sich bisher vor allem auf die Extraktion von relevanten Wörtern fokussiert. Dafür werden die Stellenanzeigen zunächst in einzelne Abschnitte zerteilt, die jeweils bestimmten Kategorien zugeordnet werden. Ein Abschnitt kann beispielsweise die Unternehmensbeschreibung, die Anforderungsbeschreibung, die Tätigkeitsbeschreibung oder Vermischtes wie die Kontaktinformationen des Unternehmens enthalten. Indem die Teile der Anzeige diesen Kategorien zugeordnet werden, kann sich die Suche der zu extrahierenden Terme einschränken lassen, denn geforderte Kompetenzen stehen zum Beispiel mit großer Wahrscheinlichkeit in dem Abschnitt zum Anforderungsprofil. Durch diese vorgeschaltete Klassifikation können Falsch Positive reduziert werden. Außerdem verringert sich der Rechenaufwand im Rahmen der Informationsextraktion, da der Umfang der zu durchsuchenden Texte reduziert wird (Hermes und Schandock 2016).

Nachdem alle Anzeigen in klassifizierte Abschnitte unterteilt sind, können in den jeweiligen Abschnitten beispielsweise sowohl neue als auch bereits bekannte Kompetenzen gesucht werden. Für die Ermittlung von potenziell relevanten Begriffen werden Bootstrapping-Verfahren angewandt, die mithilfe von bekannten Mustern neue Kompetenzen identifizieren. Jedoch wurde im Bootstrapping nur ein Durchgang durchgeführt, weil die unüberwachte Verwendung neuer Muster zu vielen falsch positiven Extraktionen geführt hat. Die initial gesetzten Muster bestehen aus einer Kombination von Lemmata und Wortarten. Wie vorherige Untersuchungen im Projekt zeigten, sind Stellenanzeigen oft ähnlich formuliert. So deutet beispielsweise das Muster

kenntnis|erfahrung + APPR + ART + NN|NE

darauf hin, dass es sich beim Nomen bzw. Eigennamen um eine Kompetenz handelt. Damit würden sowohl *Erfahrung mit dem Softwaresystem* als auch *Kenntnisse über das Supply Chain Management* als Kompetenzen aufgefunden werden. Indem solche Muster gesammelt wurden, konnten noch unbekannte Kompetenzen in den Stellenanzeigen entdeckt werden (Geduldig 2017). Bereits bekannte Kompetenzen können ebenfalls in Stellenanzeigen identifiziert werden. Die Liste der bekannten Kompetenzen setzt sich zum einen aus Kompetenzen, die in Kompetenz-Taxonomien aufgeführt werden (siehe 4.2.), und zum anderen aus den durch Bootstrapping ermittelten Kompetenzen zusammen. Das Matching von bekannten

Kompetenzen in Stellenausschreibungen soll vor allem später dazu dienen, die Stellenanzeigen mit den im Text enthaltenen Kompetenzen zu annotieren.

Sowohl die Extraktion von neuen Kompetenzen und Arbeitsmitteln als auch das Matching von bekannten Begriffen setzen Anwendung von semantischem Wissen voraus, denn die zu extrahierenden Strings sind zunächst noch nicht mit Bedeutung verknüpft. Bei neuen Ausdrücken wäre es wünschenswert, wenn diese in eine bestehende Taxonomie (siehe 4.2.) eingebunden werden könnten, denn oft handelt es sich bei diesen Strings um orthografische Varianten von bereits bekannten Kompetenzen oder Arbeitsmitteln. Beispielsweise könnte der Ausdruck *Apache-Server* bereits einen Platz in der Taxonomie eingenommen haben. Wenn jetzt allerdings der Ausdruck *Apache Server* als Kompetenz im Text erkannt wird, sollte dieser neue String mit dem bestehenden Knoten verknüpft werden, da die beiden Ausdrücke die gleiche Bedeutung haben. Im Projekt wurde diese Verknüpfung bisher teilweise über die Berechnung der String-Ähnlichkeit gelöst. Allerdings verursacht die automatische Zusammenführung von ähnlichen Strings auch viele Fehler. Ebenso wurde das Matching bisher ausschließlich auf der Zeichenebene ohne semantische Zusatzinformationen durchgeführt. Also kann es durchaus passieren, dass ein Ausdruck falsch zugeordnet wird, und zwar dann, wenn es für einen Ausdruck zwei mögliche Konzepte in der Taxonomie gibt oder wenn ein Ausdruck zwei Bedeutungen hat, von denen aber nur eine durch die Taxonomie repräsentiert ist. Diese Arbeit soll anfangen die Lücke im Extraktionsprozess zu schließen. Außerdem soll überlegt werden, wie neue Ausdrücke außer mithilfe der String-Ähnlichkeit in der Taxonomie ergänzt werden können.

4.1.2. Die Daten

Das Framework *quenfo* soll letztendlich dazu dienen, ein großes Korpus von Online-Stellenanzeigen aufzubereiten. Einmal jährlich erhält das BIBB von der Bundesagentur für Arbeit (BA) einen Auszug des Jobportals, der aus allen zum 15. Oktober aktiven Stellenanzeigen besteht. Diese wurden ursprünglich ausschließlich zum Zweck der Arbeitsvermittlung gesammelt. Eine Nutzung in der Wissenschaft ist daher nur unter strengen Auflagen, die das Urheberrecht schützen und bestimmten Datenschutzstandards entsprechen, möglich. Beispielsweise kann auf das Korpus der BA nur von einem Stand-Alone-PC ohne Internetzugang im BIBB zugegriffen werden. Durch die räumliche Distanz ist die Entwicklung des Frameworks mit diesen Daten nur erschwert möglich. Deshalb wurde für die Entwicklung der vorliegenden Arbeit auf ein anderes Korpus zugegriffen. Dieses besteht aus 10.000 Online-

Stellenanzeigen, die der Dienstleister TextKernel[5] von unterschiedlichen Job-Plattformen gescraped hat. Im Folgenden wird ausschließlich mit diesen Daten gearbeitet. Jedoch soll es nach Fertigstellung dieser Arbeit möglich sein, die Erkenntnisse und die implementierten Anwendungen innerhalb des Projektes *quenfo* zu nutzen, um sie auf die Daten der BA anzuwenden.

Die Daten von TextKernel lagen zunächst in einer .csv-Datei vor, die pro Stellenanzeige lediglich eine ID, Information über die Sprache der Anzeige sowie die Anzeige selbst bereitstellt. Es gibt also keine strukturierten Informationen über die in der Anzeige geforderten Qualifikationen. Die Stellenanzeigen sind branchenunspezifisch und kommen somit beispielsweise sowohl aus dem Bereich der IT-Dienstleistung als auch dem Gesundheitswesen. Des Weiteren befinden sich unter den Ausschreibungen Stellen für Vollzeittätigkeiten sowie für Ausbildungsplätze oder Praktika. Im folgenden Kapitel werden nun zunächst Möglichkeiten vorgestellt, wie Qualifikationen strukturiert beschrieben werden können. Darauf folgt dann ein Kapitel, das die linguistischen Eigenschaften von Stellenanzeigen näher beleuchtet.

4.2. Taxonomien zur Beschreibung von Kompetenzen

Die Messung von Kompetenzangebot und -nachfrage ist ein wichtiges Instrument der Arbeitsmarktberatung und -forschung und bedarf deshalb einer systematischen Beschreibung von Kompetenzen. Zum einen haben Arbeitnehmer so unter anderem die Möglichkeit genauere Informationen zu dem Anforderungsprofil einer Stelle zu erhalten. Zum anderen kann so auf der Nachfrageseite der Bedarf von bestimmten Qualifikationen quantitativ analysiert werden. Außerdem ist es durch die systematische Beschreibung von Kompetenzen möglich, die Ähnlichkeit von Berufen zu bestimmen. Diese Information wird beispielsweise bei der Vermittlung von beruflichen Seiteneinsteigern benötigt, um einschätzen zu können, welche alternativen Beschäftigungsmöglichkeiten für eine Erwerbsperson infrage kommen und welche Qualifikationsmaßnahmen dabei notwendig sind. Im Folgenden werden zwei bestehende Taxonomien für Kompetenzen näher beschrieben. Taxonomien bilden Klassen auf „der Grundlage von Ähnlichkeitsbeziehungen zwischen den der Taxonomie unterworfenen Gegenständen“, in diesem Fall also zwischen den Kompetenzen (Metzler Lexikon Sprache, 5. Auflage, s.v. „Taxonomie“). Dabei werden die einzelnen Konzepte zum Beispiel in hierarchische Beziehungen miteinander gesetzt. Dadurch lassen sich Aussagen darüber treffen,

[5] https://www.textkernel.com/de/ (zuletzt aufgerufen: 05.03.20)

ob eine Kompetenz eine andere verallgemeinert oder spezifischer ausdrückt. Diese Strukturen ähneln der Organisation von Kategorienkonzepten, wie sie in Kapitel 2 vorgestellt wurden.

4.2.1. Europäische Klassifikation für Fähigkeiten, Kompetenzen, Qualifikationen und Berufe (ESCO)

Eine unter vielen Kompetenz-Taxonomien ist die Europäische Klassifikation für Fähigkeiten, Kompetenzen, Qualifikationen und Berufe (englisch: *European Skills, Competence, Qualifications and Occupations*, kurz: ESCO). Sie wird von der Europäischen Kommission entwickelt und bildet verschiedene Konzepte ab, die den Arbeitsmarkt betreffen. Die ESCO ist ein multilinguales Klassifikationssystem, mit dessen Hilfe Fähigkeiten, Kompetenzen, Qualifikationen und Berufe identifiziert und kategorisiert werden können (Jones 2017, S. 10). Die ESCO unterscheidet klar zwischen Konzepten und Ausdrücken. Im Mittelpunkt stehen Konzepte, die jeweils durch verschiedene sprachliche Ausdrücke dargestellt werden können (Jones 2017, S. 11). Durch die Mehrsprachigkeit wird ferner ermöglicht beispielsweise Berufe auf mehrsprachlicher Ebene innerhalb der gesamten EU zu vergleichen. Es liegt also eine Taxonomie vor, die semantische Konzepte abbildet (siehe 2.1.). Jedes Konzept wird durch einen *Uniform Resource Identifier* (URI) repräsentiert. Dadurch wird gewährleistet, dass jeder Knoten eindeutig identifiziert werden kann. Die ESCO folgt hier den Prinzipien von *Linked Open Data*. Die Entwickler*innen möchten damit gewährleisten, dass das System möglichst einheitlich entwickelt wird und über Programmierschnittstellen angespielt werden kann. Mithilfe von standardisierten Wörterbüchern kann die Taxonomie erweitert werden. Dafür verwenden die Entwickler*innen verschiedene Formate wie SKOS-RDF oder CSV, außerdem lässt sich die Taxonomie über eine lokale API sowie eine Web Service API abfragen. Dadurch kann die Taxonomie in andere Anwendungen eingebaut werden (Jones 2017, S. 35). So soll die ESCO dafür genutzt werden, Tools im Bereich des Job Matchings oder der Karriereberatung zu unterstützen. Außerdem kann die Bevölkerung sich über ein Portal[6] zu den Anforderungen auf dem Arbeitsmarkt informieren.

Für die vorliegende Arbeit sind nur Konzepte aus dem Bereich der Kompetenzen und Fähigkeiten relevant, daher wird im Folgenden nur dieser Bereich beschrieben. In der ESCO sind Fähigkeiten, Kenntnisse und Kompetenzen in einer als „Skills“ bezeichneten Säule zusammengefasst. Fähigkeiten (*Skill*) definieren die Entwickler*innen der ESCO als die Befähigung Wissen anzuwenden und zu nutzen, um Aufgaben abzuschließen und Probleme zu lösen. Dabei können Fähigkeiten kognitiver oder praktischer Natur sein. Kenntnisse

[6] https://ec.europa.eu/esco/portal (zuletzt aufgerufen: 17.03.20)

(*Knowledge*) hingegen beziehen sich auf Fakten, Prinzipien, Theorien und Praktiken, die für ein spezifisches Arbeits- oder Forschungsfeld relevant sind, und das Ergebnis eines Lernprozesses darstellen. Sobald sowohl Kenntnisse als auch Fähigkeiten zusammen mit sozialem und methodologischem Können eingesetzt werden, um Arbeits- oder Forschungssituationen zu meistern, spricht man von einer Kompetenz (*Competence*)(Jones 2017, S. 19). In der ESCO wird in der Datenstruktur keine Unterscheidung zwischen Fähigkeiten und Kompetenzen gemacht (Jones 2017, S. 20). Da die Taxonomie mehrsprachig abgebildet ist, gibt es für jedes Konzept mindestens einen sprachlichen Ausdruck pro Sprache. Synonym verwendete Ausdrücke werden unter einem Knoten zusammengefasst. Dabei wird allerdings immer ein bevorzugter Ausdruck ausgewählt, der als Hauptausdruck in der jeweiligen Sprache fungiert. Alle Hauptausdrücke sind innerhalb einer Sprache in der gesamten Taxonomie einzigartig, um Ambiguitäten zu vermeiden (Jones 2017, S. 12).

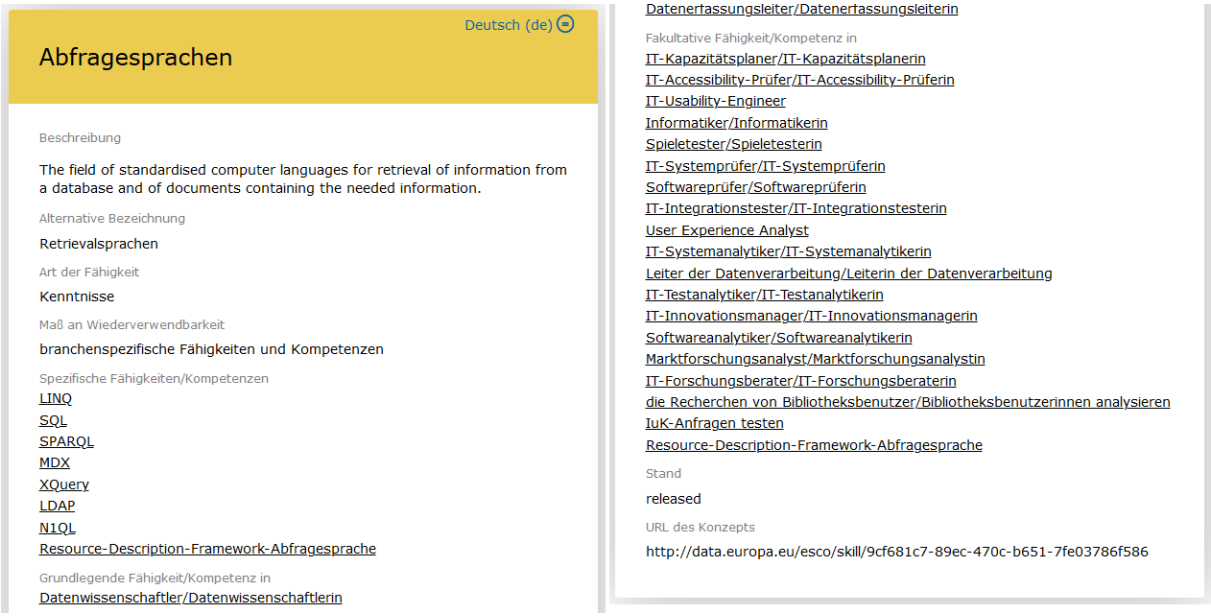

Abbildung 5: Beispiel-Eintrag ESCO

Im Online-Portal der ESCO[7] können alle Konzepte inklusive sämtlicher Informationen dazu eingesehen werden (siehe Abbildung 5). Jeder Eintrag besteht dabei aus einem Hauptausdruck pro Sprache (im Beispiel: „Abfragesprachen“), einer Beschreibung in englischer Sprache, einem Hinweis, ob es sich um eine Kompetenz oder eine Kenntnis handelt, das Maß an Wiederverwendbarkeit sowie der URI des Konzepts. Außerdem kann das Konzept mit spezifischeren oder allgemeineren Knoten verknüpft sein. Im hier vorgestellten Fall wird zum Beispiel XQuery als Hyponym angegeben. Ferner folgt eine Auflistung, in welchen Berufen die Fähigkeit zwingend verlangt wird oder optional vorhanden sein sollte. Je nach Eintrag

[7] https://ec.europa.eu/esco/portal/skill (zuletzt aufgerufen: 13.02.20)

können auch alternative Bezeichnungen ausgelistet sein. Im Beispiel wird „Retrievalsprachen“ als Synonym für „Abfragesprachen“ genannt. Durch diese Struktur kommen die Einträge in der ESCO den in Kapitel 2 beschriebenen semantischen Konzepten relativ nah. In beiden Fällen kann es mehrere Ausdrücke geben, die auf das Konzept bzw. den Eintrag referenzieren. Und während Konzepte hierarchisch organisiert sind, können Einträge in der ESCO mit allgemeineren Kompetenzen verknüpft werden.

Für diese Arbeit wurde auf die ESCO v. 1.0.3 zurückgegriffen, die nun genauer untersucht werden soll[8]. Sie umfasst 13.485 Skill-Konzepte, die teilweise mehrere Ausdrücke pro Konzept enthalten. Insgesamt gibt es 18.822 verschiedene Ausdrücke, davon besteht gut ein Drittel aus zwei Tokens (siehe Tabelle 1). Diese Bigramme enthalten häufig ein Verb, es wird also meist eine Tätigkeit beschrieben, wie zum Beispiel bei *Karriereberatung anbieten*, *Kaffeebohnen bewerten* oder *Paletten laden*. Ausdrücke, die aus ein oder zwei Tokens bestehen, können in den lemmatisierten Stellenanzeigen relativ schnell über ein einfaches String-Matching identifiziert werden.

Tabelle 1: Anzahl der Tokens pro Skill-Ausdruck

	1	2	3	4	5	6	7	8	9 oder mehr
ESCO	2.524	6.963	2.965	2.845	1.545	971	472	265	272
AMS	14.681	4.961	4.842	1.609	776	241	112	37	32

Im Vergleich dazu findet sich unter den besonders langen Ausdrücken mit 9 oder mehr Tokens beispielsweise folgende Formulierung: *Grundregeln der Pflege und Wartung für Lederwaren und Maschinen für Schuhwerk anwenden*. Für diese Ausdrücke ist es deutlich unwahrscheinlicher, dass man genau diese Formulierung in Stellenanzeigen findet.

4.2.2. Das Qualifikations-Barometer des Arbeitsmarktservice Österreich (AMS)

In Österreich gibt es ebenfalls ein Instrument zur systematischen Beschreibung von Qualifikationen auf dem Arbeitsmarkt, allerdings wird dieses auf nationaler Ebene entwickelt. Der Arbeitsmarktservice Österreich (AMS) entwickelt seit 2002 das sogenannte Qualifikations-Barometer[9], mit dessen Hilfe ähnliche Zwecke wie mit der ESCO verfolgt werden. Zum einen soll eine Plattform entstehen, die Bürger*innen Informationen zu Trends in der beruflichen Qualifikation bereitstellt. Zum anderen werden Methoden entwickelt, die das Monitoring des

[8] exploration/taxonomy_exploration.py
[9] http://bis.ams.or.at/qualibarometer/kompetenzstruktur.php (zuletzt aufgerufen: 05.03.20)

Arbeitsmarkts erleichtern sollen. Vor der Entwicklung des Qualifikations-Barometers wurde diese Aufgabe von vereinzelten Studien in Österreich übernommen, die den Bedarf an beruflichen Kompetenzen im Land ermitteln sollten. Da die Studien oft auf politischer Ebene initiiert wurden, schwankten das Studiendesign und die Dauer der Durchführung je nach aktueller Notwendigkeit. Beispielsweise wurden um die Jahrtausendwende viele Studien zu IT-Kenntnissen durchgeführt, um den ansteigenden Bedarf von Fachkräften in diesem Bereich zu prognostizieren (Markowitsch et al. 2007, S. 192). Damit die Forschung in diesem Bereich strukturierter durchgeführt werden konnte, wurden 2002 zwei Unternehmensberatungen vom AMS damit beauftragt, ein System zu entwickeln, das die kontinuierliche Beobachtung des Kompetenzbedarfs auf dem Arbeitsmarkt zulässt (Markowitsch et al. 2007, S. 195).

Genau wie die ESCO enthält das Qualifikations-Barometer sowohl eine Taxonomie für Berufe als auch eine Taxonomie für Fähigkeiten und Kompetenzen. Diese werden kontinuierlich aktualisiert, um alle den Arbeitsmarkt betreffenden Konzepte abzudecken (Markowitsch et al. 2007, S. 198). Für diese Arbeit wurde ein Ausschnitt[10] des Qualifikations-Barometers gewählt, der innerhalb von *quenfo* vor einigen Jahren gescraped wurde. In dem Ausschnitt befinden sich 27.291 verschiedene Ausdrücke. Der überwiegende Anteil von 14.681 Ausdrücken besteht aus Unigrammen (siehe Tabelle 1). Die Bigramme enthalten weniger Verben als die Bigramme in der ESCO-Taxonomie. Im Qualifikations-Barometer sind die Kompetenz-Ausdrücke außerdem mit 1,9 Tokens pro Ausdruck kürzer als die in der ESCO-Taxonomie mit durchschnittlich 3,1 Tokens pro Ausdruck.

Sowohl im Qualifikations-Barometer als auch in der ESCO enthalten einige Knoten bereits orthografische Varianten. Jedoch enthalten Stellenanzeigen oft weiterhin neue Varianten, mit denen die Taxonomien angereichert werden sollten. Eine Erweiterung der ESCO könnten beispielsweise innerhalb des Projekts durch weitere Tupel im SKOS-RDF geschehen. Im Gegensatz dazu gestaltet sich die eigenmächtige Erweiterung des Qualifikations-Barometers vom AMS schwierig, weil hier neben dem Portal keine Dateien veröffentlicht werden.

4.3. Sprachliche Eigenschaften von Stellenanzeigen

Für die computerlinguistische Verarbeitung der Stellenausschreibungen wurden innerhalb des Projekts *quenfo* bisher allgemeinsprachliche Modelle und Bibliotheken verwendet. Diese Entscheidung begründet sich darin, dass die Ergebnisse von Modellen, welche beispielsweise auf Artikeln der Wikipedia trainiert wurden, ausreichten, um die Anzeigen mithilfe von *quenfo*

[10] siehe data/skills/AMS_CategorizedCompetences.db

weiter zu verarbeiten. Trotzdem lohnt es sich, Stellenanzeigen auf ihre sprachlichen Eigenschaften hin zu untersuchen, um in Zukunft möglicherweise NLP-Modelle speziell für Stellenanzeigen entwickeln zu können.

Äußerlich fällt vor allem auf, dass in Stellenanzeigen viele Aufzählungen auftreten (siehe Abbildung 6). Das unterscheidet sie von anderen Forschungsgegenständen im Text Mining wie beispielsweise Romanen oder Zeitungsartikeln. Weiter sind Stellenanzeigen oft in einzelne Abschnitte unterteilt, die sich jeweils auf unterschiedliche Inhalte wie beispielsweise das Anforderungsprofil oder die Unternehmensbeschreibung beziehen. Unter 3.1.1. wurde bereits gezeigt, dass diese Unterteilungen bereits im Projekt genutzt wurden, um den Suchbereich für relevante Informationen einzugrenzen. Außerdem ist anzumerken, dass es sich bei dem hier untersuchten Genre um keine lektorierten Texte handelt. Zwar werden die Stellenausschreibungen vor allem in größeren Unternehmen vermutlich durch ein Korrekturverfahren geschickt, jedoch enthalten Stellenanzeigen trotzdem oft Rechtschreibfehler und vor allem kein standardisiertes Vokabular. Dazu kommt, dass in einigen Anzeigen mit spezieller Groß- und Kleinschreibung gearbeitet wird, zum Beispiel indem sämtliche Buchstaben groß bzw. klein geschrieben werden. Diese Tatsachen wirken sich auch auf die Güte der Ergebnisse der NLP-Tools bei der linguistischen Vorverarbeitung aus.

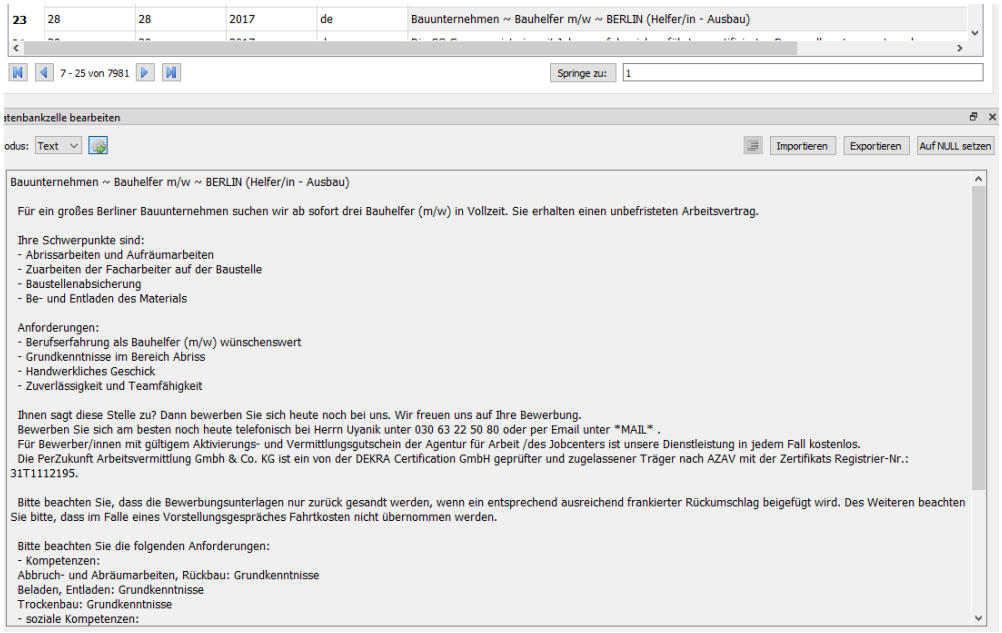

Abbildung 6: Beispiel für eine Stellenanzeige

In Bezug auf die Aufzählungen kann außerdem angemerkt werden, dass die einzelnen Punkte häufig keinen vollständigen Sätzen entsprechen, was sich auch auf die Verteilung der Wortarten im Stellenanzeigen-Korpus auswirken könnte. Um diese Vermutung zu überprüfen, wurde zum

Vergleich ein Auszug aus dem *Corpus of German Language Fiction* (CGLF) herangezogen (Fischer und Strötgen 2017). Damit ein möglicher syntaktischer Wandel über die Zeit möglichst wenig Einfluss auf das Ergebnis hat, wurden nur Texte, die später als 1900 verfasst wurden, ausgewählt. Vergleicht man die Verteilung der Wortarten innerhalb der beiden Korpora miteinander[11], fällt vor allem auf, dass Nomina in Stellenanzeigen besonders überrepräsentiert sind, während Verben, Adverbien und Pronomen deutlich seltener als in den Prosa-Texten auftreten (siehe Abbildung 7). Auch Adjektive tauchen häufiger in den Stellenanzeigen als in den Prosatexten auf. Das könnte vor allem daran liegen, dass die Beschreibungen von Soft-Skills in Stellenanzeigen größtenteils aus Adjektiven bestehen.

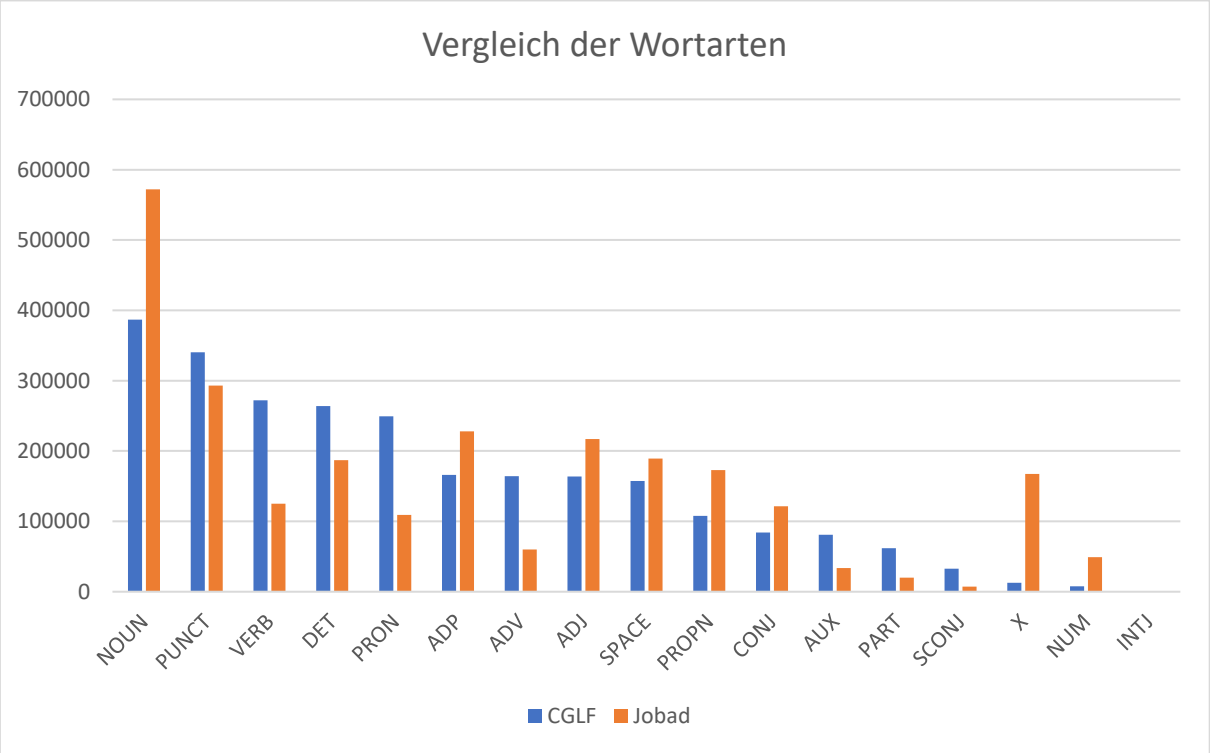

Abbildung 7: Vergleich der Wortarten (je Korpus ca. 2.5 Mio. Tokens)

Des Weiteren sind im Stellenanzeigen-Korpus auffällig viele Wörter mit dem X-Tag markiert. Das lässt zunächst vermuten, dass das im POS-Tagging verwendete Modell nicht optimal auf die Sprache in Stellenanzeigen anwendbar ist. Eine erste Sichtung der Wörter in dieser Kategorie zeigt, dass vor allem englischsprachige Wörter sowie Teile von koordinierten Ausdrücken entsprechend markiert wurden. Da Morphemkoordinationen wie beispielsweise *Bau- und Landmaschinen* häufig in Stellenanzeigen verwendet werden, erklärt dies teilweise die Größer der X-Kategorie. Jedoch ist oft auch die fehlende Großschreibung Ursache für eine Fehlzuweisung. Um diesen Problemen entgegenzuwirken, könnte man einerseits ein neues

[11] siehe exploration/pos_comparison.py

Modell speziell auf annotierten Stellenanzeigen trainieren oder andererseits die Texte der Anzeigen vor dem POS-Tagging entsprechend bearbeiten, sodass beispielsweise englischsprachige Teile entfernt werden, da diese die computerlinguistische Bearbeitung bisher nicht unterstützen können. Wenn man die fehlerhaften Zuweisungen allerdings zunächst ignoriert, lassen sich trotzdem durch diesen Vergleich quantitativ erste Beobachtungen machen. Zum Beispiel wird deutlich, dass Stellenanzeigen eine Domäne bilden, deren Eigenschaften auch auf die Syntax der Texte abfärben. Um diese Vermutung jedoch zu festigen, sollte man Korpora aus weiteren Domänen heranziehen, um die Wortarten der Stellenanzeigen mit diesen erneut gegenüberzustellen.

Ein Korpus kann ebenfalls durch die quantitativen Eigenschaften des Wortschatzes beschrieben werden. Dafür kann unter anderem das *Type-Token-Ratio* (TTR) genutzt werden, das wie folgt definiert ist:

$$TTR(Corpus) = \frac{Types(Corpus)}{Tokens(Corpus)}$$

Das TTR bildet das relative Verhältnis zwischen der Menge der Types und der Menge der Tokens ab und kann somit den lexikalischen Reichtum einer Textsammlung unabhängig von der Größe des Korpus angeben. Oft wird dieser Wert immer jeweils über 1.000 Tokens errechnet, um anschließend daraus letztendlich das Mittel zu bilden. Das TTR für das Stellenanzeigen-Korpus beträgt 0.52[12]. Das bedeutet, dass jedes Type durchschnittlich nur zweimal in 1.000 Wörtern enthalten ist. Der Ausschnitt des CGLF ist im Vergleich dazu mit einem TTR von 0.4 weniger abwechslungsreich. Diese Beobachtung überrascht, da die intuitive Annahme in diesem Punkt sein könnte, dass Stellenanzeigen lexikalisch eher einseitig sind. Diese Annahme begründete sich darin, dass Stellenanzeigen immer nach ähnlichen Mustern aufgebaut sind und demnach oft die gleichen Ausdrücke verwenden müssten (siehe 4.1.1.). Jedoch muss man bei der Bildung des TTR beachten, dass jede Anzeige im Korpus eine andere Anstellung bewirbt und dementsprechend unterschiedliche Ausdrücke für die Kompetenzen, Tätigkeiten und Arbeitsmittel enthält. Im Gegensatz dazu ist die Handlung der Texte aus dem anderen Korpus vermutlich pro Text einheitlich, sodass sich ein niedrigeres TTR ergibt. In den Stellenanzeigen könnten auch vor allem Eigennamen dafür sorgen, dass das TTR relativ hoch ausfällt, da diese vermutlich nur selten doppelt verwendet werden. Um dies zu überprüfen, könnte man das TTR für alle offenen Wortartenklassen separat berechnen.

[12] siehe exploration/type_token_ratio.py

Speziell in Bezug auf die distributionellen Eigenschaften von Wörtern in Stellenanzeigen und die daraus resultierende Bedeutung wird abschließend ein Word-Embedding[13], das aus Wikipedia-Artikeln und Zeitungsartikeln erzeugt wurde, mit einem aus den Stellenanzeigen erzeugten Embedding verglichen. Eine Hypothese, die für domänenspezifische Bedeutungen spricht, ist, dass sich die Verteilung der Wörter in den Embeddings unterschiedlich verhält. Beide Modelle wurden mit den gleichen Methoden erzeugt und anschließend auf die Schnittmenge ihrer Wörterbücher reduziert[14]. Dadurch befinden sich anschließend in beiden Modellen 2419 Vektoren für ein identisches Wörterbuch. Die räumliche Verteilung der Vektoren wird anhand der nächsten Nachbarn im jeweiligen Vektorraum verglichen[15]. Dabei interessieren vor allem die Vektoren, die in beiden Modellen über ein ähnliches Umfeld verfügen, sowie die Vektoren, deren Nachbarn sehr unterschiedlich ausfallen. Wir betrachten zunächst die zehn nächsten Nachbarn eines Vektors. In der quantitativen Analyse fällt vor allem auf, dass sich die Nachbarschaften von einem Ausdruck in den zwei verschiedenen Modellen nur wenig überschneiden. Bei 1739 Wortvektoren stimmen nur weniger als drei nächste Nachbarn überein, davon teilen 577 Vektoren überhaupt keine gleichen Nachbarn (siehe Tabelle 2). Im Gegensatz dazu stimmen lediglich bei 45 Wörtern mindestens 7 der 10 nächsten Nachbarn überein. Unter diesen Wort-Vektoren sind vor allem Zahlen-Wörter, Städtenamen, aber auch Nomen wie *Deutsch*, *Mathematik*, *Master* oder *Studiengang*.

	0 gleiche Nachbarn	*1-29% gleiche Nachbarn*	*30-69% gleiche Nachbarn*	*70-100% gleiche Nachbarn*
Bei 10 Nachbarn	577	1162	635	45
Bei 15 Nachbarn	358	1565	478	18
Bei 20 Nachbarn	244	1570	586	19

Tabelle 2: Anzahl der übereinstimmenden Nachbarn

[13] https://devmount.github.io/GermanWordEmbeddings/ (zuletzt aufgerufen: 24.03.20)
[14] siehe preprocessing/model_intersection.py
[15] siehe exploration/embedding_comparison.py

Im Vergleich dazu findet man in der Menge der Vektoren ohne gemeinsame Nachbarn Wörter wie *Schule*, *Anforderungen*, *Personal* oder *Arbeit*. Während sich im Modell der Stellenanzeigen im Umfeld von *Schule* zum Beispiel Vektoren zu Wörtern wie *Stiftung*, *Wirtschaft*, *Hochschule* oder *Technik* befinden, zeichnen sich im anderen Modell Wörter wie *Kita*, *Klasse*, *Abitur* oder *Ausbildung* also nächste Vektor-Nachbarn ab. Diese Unterschiede deuten an, dass sich die Bedeutung des Wortes *Schule* zwischen den Domänen verschiebt. Während das Wort in Wikipedia- und Zeitungsartikeln anscheinend vermehrt in Bezug auf die Förderung von Minderjährigen auftaucht, handelt es sich in den Stellenanzeigen vermutlich um Kontexte, die sich mit fortgeschritteneren Bildungsniveaus, wie zum Beispiel der Ausbildung befassen. Diese Hypothese müsste jedoch noch weiter überprüft werden, zum Beispiel, indem ein weiteres Modell auf der Grundlage anderer Daten zum Vergleich herangezogen würde. Die Tendenz, dass die meisten Wörter eher zu einer geringen Überschneidung neigen, setzt sich ebenfalls fort, wenn man den Rahmen der nächsten Vektoren auf 15 bzw. 20 Nachbarn erweitert. Dabei verringern sich gleichzeitig die Anzahl der Vektoren mit einer stark übereinstimmenden Nachbarschaft. Jedoch kann man dieses Phänomen auch damit erklären, dass die Wahrscheinlichkeit, dass viele Elemente einer Gruppe gleich sind, mit einer Vergrößerung der Gruppe sinkt. Außerdem sind in unserem Fall weiter entfernte Nachbarn vermutlich nicht mehr so aussagekräftig für ein Wort wie nähere Nachbarn.

Die bisherigen Untersuchungen haben gezeigt, dass Stellenanzeigen sowohl in Bezug auf die Syntax als auch auf die Semantik eine eigene Domäne bilden. Außerdem wurden bestehende Kompetenz-Taxonomien sowie das bisherige Framework zur Informationsextraktion aus Stellenausschreibungen vorgestellt. Aufbauend auf diese Informationen soll in den folgenden Kapiteln nun untersucht werden, welche Methoden dabei helfen, extrahierte Kompetenzen einerseits zu disambiguieren und andererseits neu in eine bestehende Taxonomie einzuordnen.

5. Disambiguierung

Die vorherigen Kapitel zeigen zum einen wie semantische Konzepte und Text zusammenhängen und beleuchten zum anderen Stellenanzeigen als Forschungsgegenstand. Dieses Wissen soll nun exemplarisch genutzt werden, um Ausdrücke, die im Zuge der Informationsextraktion als potenzielle Kompetenzen markiert wurden, zu untersuchen, und festzustellen, ob diese korrekt extrahiert wurden. Dabei gilt es einerseits festzustellen, ob der Ausdruck in dem jeweiligen Kontext eine Kompetenz darstellt, und andererseits die Zuordnung in die Taxonomie zu überprüfen. Letzteres ist vor allem aus Sicht der klassischen Disambiguierung interessant, weil es hier um die Zuordnung zwischen Ausdrücken im Text und semantischen Konzepten in der Taxonomie geht. Allerdings zeigte ein erster Blick auf die aus dem Korpus extrahierten Ausdrücke, dass sich unter den Extraktionen kaum Zuordnungen befinden, die im Hinblick auf diesen Aspekt falsch zugeordnet wurden. Dadurch ist es nicht möglich, ein Disambiguierungssystem, wie es unter 3.1. beschrieben wurde, mit den Stellenanzeigen quantitativ zu evaluieren. Was im Zuge der Informationsextraktion allerdings häufig auftritt, sind Falschzuordnungen, die dem ersten Schema entsprechen. Hier fällt vor allem auf, dass Wörter, die beispielsweise in der ESCO auftauchen, in einigen Kontexten der Stellenausschreibungen nicht als Bewerber*innen-Kompetenz gemeint sind. Diese falschen Extraktionen gilt es nun in einem nachgeschalteten Verfahren herauszufiltern. Genau wie Calanca et al. in ihrer Analyse zu Soft Skills feststellten, gibt es einige Begriffe, die je nach Kontext auf einen Soft Skill oder etwas anderes – und damit Irrelevantes – referenzieren. Jedoch beschränkt sich dieses Problem nicht nur auf Soft Skills, sondern auf sämtliche Gruppen der Kompetenzen. So ist zum Beispiel im Satz *Sie beherrschen die **deutsche** Sprache in Wort und Schrift.* die Kompetenz vollkommen richtig extrahiert worden, während für den Satz *Für ein **deutsches** Unternehmen suchen wir neue Mitarbeiter*. eine falsche Extraktion vorgenommen wurde. Denn hier bezieht sich das Adjektiv *deutsch* nicht auf die Kompetenz des Bewerbers bzw. der Bewerberin, sondern beschreibt eine Eigenschaft des Unternehmens. Diese Tatsache führt im Rahmen der Informationsextraktion zu vielen Falsch Positiven, die durch eine nachgeschaltete Filterung verringert werden sollen. Im Folgenden wird nun beschreiben auf welcher Datengrundlage diese Filterung durchgeführt wurde und wie die Disambiguierung konkret implementiert und evaluiert wurde.

5.1. Referenzdaten & Vorverarbeitung

Für die Evaluation eines überwachten Ansatzes zur Disambiguierung wurden zunächst Trainingsdaten erstellt, indem der Output des Matching-Programms von *quenfo* manuell

validiert wurde. Dabei wurde für jeden Satz überprüft, ob dort eine Kompetenz richtig identifiziert wurde oder nicht. Dies wurde entsprechend in einer Spalte der Output-Datenbank[16] mit 0 oder 1 annotiert. Die Annotation bezieht sich hier ausschließlich auf eine falsche Extraktion und nicht auf eine falsche Zuordnung innerhalb der Taxonomie. In der Datenbank befinden sich 1.071 Einträge, die korrekt extrahiert wurden und 779 Einträge, bei denen eine falsche Extraktion durchgeführt wurde. Weitere 20.208 Einträge sind noch nicht annotiert und könnten deshalb in zukünftigen Untersuchungen beispielsweise für unüberwachte Verfahren verwendet werden (Agirre et al. 2006). Die Kompetenzen, die annotiert wurden, sind inhaltlich ganz unterschiedlich und reichen dabei von Wörtern wie *Personalberatung* über *zeichnen* bis hin zu *Freundlichkeit*. Sie sind semantisch also nicht verwandt und sollten deshalb auch über unterschiedliche Kontexte verfügen. Außerdem schwankt der Anteil der richtig bzw. falsch extrahierten Einheiten pro Kompetenz. Im Fall von *deutsch* sind beispielsweise 274 Extraktionen tatsächlich korrekt und 73 Einträge falsch.

5.2. Aufgabenstellung

Die hier verwendeten Methoden zur Disambiguierung lehnen sich an den unter 3.1. vorgestellten Ansatz von Resnik an, bei dem Ausdrücke aufgrund ihrer Selektionspräferenzen mit Bedeutung verknüpft wurden. In Bezug auf die potenziellen Kompetenzen in Stellenanzeigen interessiert uns, ob die Untersuchung der Selektionsbeschränkungen ebenfalls dabei helfen kann, falsch positive Extraktionen zu minimieren. Dafür wird für jeden Ausdruck, der vom IE-System extrahiert wurde, der Valenzträger im Dependenzbaum identifiziert. Diese Information wird anschließend genutzt, um zu entscheiden, ob die Extraktion richtig war oder ob sich der Ausdruck auf keine Kompetenz bezieht. Wie bereits unter 2.2 erläutert, kann Disambiguierung auch als Klassifikationsproblem aufgefasst werden. Klassifikationsalgorithmen, die solche Probleme lösen, bestehen oft aus einem Klassifikationsmodell, welches so auf das Zuweisungsproblem zugeschnitten ist, dass es Datenobjekte möglichst fehlerfrei in Kategorien einordnet. Je nach Algorithmus-Typ kann das Modell unterschiedlich aufgebaut sein. Beispielsweise kann es aus verschachtelten Klassifikationsregeln nach dem Wenn-Dann-Prinzip bestehen, die die Datenobjekte auf bestimmte Eigenschaften hin überprüfen und sie daraufhin den Kategorien zuweisen. In unserem Fall der Disambiguierung ist das Ziel also ein Modell zu finden, das Sätze, in denen potenzielle Kompetenzen auftauchen, dahingehend klassifiziert, ob sie tatsächlich eine Kompetenz enthalten oder nicht. In vielen Fällen wird dieses Modell, also beispielsweise die

[16] siehe data/jobads/classified_sentences.db

Klassifikationsregeln, auf der Grundlage von bereits bekannten Zuordnungen erstellt und angepasst. Diese sogenannten Trainingsdaten können beispielsweise entstehen, indem die gewünschten Kategorien zuvor manuell annotiert werden (Han et al. 2012, S. 18). In unserem Beispiel liegen bereits einige Sätze vor, die mit 0 oder 1 markiert sind (siehe 5.1.). Aus dieser Information und einigen ausgewählten Eigenschaften der entsprechenden Sätze kann ein Modell entwickelt werden, das in der Lage ist, noch unbekannte Sätze zu klassifizieren. Mit diesen Eigenschaften kann beispielsweise der Valenzträger der Kompetenz gemeint sein wie es soeben beschrieben wurde. Die genaue Implementierung wird im Folgenden beschrieben.

5.3. Software-Umsetzung

Die Klassifikation der Sätze wird durch einen Entscheidungsbaum umgesetzt. Diese bestehen aus Klassifikationsregeln, die in einer Baumstruktur verknüpft werden. Die Baumstruktur legt dabei fest, in welcher Reihenfolge die Regeln auf das zu klassifizierende Objekt angewandt werden sollen. Beim Training wird dafür geprüft, welche Regeln die Menge der bekannten Objekte im jeweiligen Schritt möglichst sauber in die bekannten Klassen unterteilt. Nachdem die Objekte nach der ersten Regel in Gruppen eingeteilt wurden, wird auf diese Gruppen jeweils wieder eine andere Regel angewandt, die diese Gruppe wiederum unterteilt. An den Enden des Entscheidungsbaums befinden sich schließlich die Blätter, die die Kategorien repräsentieren. Der abgewandelte Ansatz von Resnik wurde auf die unter 5.1. beschriebenen annotierten Sätze angewandt. Dafür wurde in Python mithilfe von NLTK ein binärer Entscheidungsbaum-Algorithmus implementiert, der vorhersagt, ob ein Ausdruck, der von der Informationsextraktion als Kompetenz identifiziert wurde, auch eine Kompetenz ist.

Für die Klassifikation der Extraktionen wurden zwei verschiedene Merkmalsauswahlen einander gegenübergestellt und verglichen[17]: Zum einen wurden Sätze anhand der potenziellen Kompetenz in lemmatisierter Form und ihres Valenzträgers klassifiziert. Auf der anderen Seite wurden die vorangehenden und nachfolgenden Tokens (ebenfalls in lemmatisierter Form) berücksichtigt. Dadurch wird die Disambiguierung sowohl mithilfe von semantisch tiefgreifenderen Informationen als auch durch einfache kontextuelle Informationen durchgeführt.

5.4. Evaluation

Um zu ermitteln, auf Grundlage welcher Merkmale ein besserer Entscheidungsbaum erstellt werden kann, wurde die Klassifikation mit jeder der Konfigurationen durchgeführt. Dafür

[17] siehe classification/categorial_classifier.py

wurde jeweils eine Matrix erstellt, die bei der anschließenden Kreuzvalidierung als Datengrundlage dient. Kreuzvalidierung ist ein Verfahren, das mehrfach aus einer Menge von Trainingsdaten Modelle erstellt, um jeweils unterschiedliche Daten zu klassifizieren. Ziel dabei ist es, dass jedes Trainingsobjekt genau einmal klassifiziert wird, sodass keine Elemente gleichzeitig sowohl in der Trainings- als auch in der Testmenge enthalten sind. Durch die Rotation der Trainings- bzw. Testdaten wird ungleichverteilte Daten vorgebeugt, die das Evaluationsergebnis verfälschen können. Im nun beschriebenen Experiment wurde eine *5-fold* Kreuzvalidierung gewählt, bei der in fünf Schritten nacheinander trainiert und evaluiert wird.

5.4.1. sequenzielle Merkmalsauswahl

Für den sequenziellen Ansatz der Merkmalsauswahl, bei dem ein Satz über die vorhergehenden und die nachfolgenden Wörter der potenziellen Kompetenz klassifiziert wird, werden nur inhaltliche Wörter aus dem unmittelbar umliegenden Kontext in die Auswahl der Merkmale aufgenommen. Für den Satz *Du bist **außerdem belastbar** und flexibel in **Zeiten hohen** Arbeitsaufkommens*, bei dem *flexibel* als potenzielle Kompetenz markiert wurde, besteht der Merkmalvektor beispielsweise aus den als fett markierten Wörtern. Auf dieser Grundlage erreicht die Klassifikation über Entscheidungsbäume eine Precision von 0.77 und einen Recall von 0.89. Falsche Kategorienzuweisungen wollen wir nun im Detail betrachten. Beispiele für False Positives sind:

1. In interdisziplinären Teams arbeiten wir mit modernen Methoden wie Design Thinking oder **Scrum** und schaffen Freiräume für eigene Ideen.
2. Bereitstellung von Daten für die **Abrechnung** und Rechnungsstellung.
3. Sie **zeichnen** sich durch eine präzise und eigenständige Arbeitsweise aus.

In Beispiel 1 und 2 bezieht sich der extrahierte Ausdruck eher auf eine Tätigkeit als auf eine Kompetenz, die bereits mitgebracht werden soll. In Beispiel 3 ist die Extraktion falsch, weil das Verb *auszeichnen* in keinem Zusammenhang mit *zeichnen* als Tätigkeit oder Fähigkeit steht. Im Gegensatz dazu finden sich unter den falsch negativ klassifizierten Sätzen unter anderem folgende Beispiele:

1. Du behältst den Überblick über die Vielzahl an unterschiedlichen Themen und kannst gut **improvisieren**.
2. Sie bringen sehr gute MS Office Kenntnisse mit (insbesondere **Excel**; VBA von Vorteil)
3. **Englische** oder weitere Sprachkenntnisse sind vorteilhaft.

Bei diesen Fällen lässt sich bisher kein richtiges Muster erkennen. Eine erste Gemeinsamkeit liegt jedoch dadurch vor, dass die potenzielle Kompetenz häufig als erstes bzw. letztes Inhaltswort in einem Satz auftritt. Dadurch bleibt der Merkmalsvektor teilweise unausgefüllt.

5.4.2. dependenzielle Merkmalsauswahl

Im Gegensatz zu den nächsten Nachbarn als Merkmale wurde außerdem die Klassifikation mithilfe der jeweiligen Valenzträger im Dependenzbaum getestet. Dabei wurden als Merkmale die lemmatisierte potenzielle Kompetenz sowie der lemmatisierte inhalttragende Kopf der Kompetenz gewählt. Letzteres meint, dass der Dependenzbaum bei der Auswahl des Kopfes so lange abgeschritten wurde, bis ein Token aus einer offenen Wortklasse oder die Wurzel des Baumes erreicht wurde. Anhand dieser zwei Merkmale erreicht die Klassifikation eine Precision von 0.86 und einen Recall von 0.93. Auch hier wollen wir kurz überprüfen, welche Falschzuweisungen durch die Klassifikation entstanden sind. Beispielsweise wurden folgende Ausdrücke[18] fälschlicherweise in die Kategorie der Kompetenzen eingeordnet:

1. **Flexible** *Arbeitszeiten* und eine gute ÖPNV-Anbindung sind zudem gewährleistet.
2. **Freundlichkeit**, Offenheit und Fairness *sind* die Grundwerte nach denen dabei gehandelt wird.
3. Die vollständigen Bewerbungsunterlagen sind in **deutscher** *Sprache* einzureichen.

Die als Kompetenz markierten Begriffe beziehen sich hier oft auf die Arbeitssituation, wie beispielsweise in Satz 1 und Satz 2. Das dritte Beispiel stellt einen Sonderfall dar, weil die Formulierung *deutsche Sprache* in vielen Fällen auf eine Kompetenz anspielt. In diesem Fall wird der Ausdruck jedoch im Kontext der Bewerbungsunterlagen erwähnt. Es geht also nicht um die Anstellung an sich, sondern den Bewerbungsprozess, selbst wenn das Verfassen der Unterlagen auf Deutsch die Sprachkenntnisse voraussetzt. False Negatives sind beim dependenziellen Ansatz unter anderem bei folgenden Sätzen aufgetreten:

1. Durch Ihre bisherigen Tätigkeiten können Sie auf ein ausgeprägtes Know-how speziell im *Bereich* **Personalbeschaffung** zurückgreifen.
2. Sie ***telefonieren*** gerne und arbeiten gerne im Kundenservice.
3. Arbeitswillen und **Freundlichkeit** *wird* vorausgesetzt.

Bei vielen Sätzen wie in Beispiel 1 und 3 ist eine tiefere Analyse der falschen Klassifizierung notwendig, da nicht sofort ersichtlich ist, was den Fehler auslöst. Des Weiteren wird anhand von Beispiel 2 deutlich, dass der dependenzielle Ansatz, wie er bisher ausgeführt wird, zu kurz

[18] Kompetenzen sind hier fett und die jeweiligen Valenzträger kursiv markiert.

greift. Denn es wird nicht berücksichtigt, was geschieht, wenn der zu untersuchende Ausdruck gleichzeitig die Wurzel im Satz ist. An dieser Stelle könnte es zum Beispiel sinnvoll sein, neben dem Elternknoten auch alle Satelliten eines Ausdrucks als Merkmale in den Vektor mit aufzunehmen. Außerdem könnte ein nächster Schritt darin bestehen, die Valenzträger bzw. die Satelliten der Kompetenzen mit taxonomiellem Wissen anzureichern.

6. Synonyme identifizieren

Unter den extrahierten Kompetenzen können genauso Ausdrücke sein, für die es noch keine Zuordnung in der Taxonomie gibt. Für die Weiterentwicklung im Projekt wäre es aber neben der Disambiguierung ebenso hilfreich, diese Terme zu kategorisieren. Wie bereits unter 4.1.1. erwähnt, wurde bisher im Projekt erprobt, ob eine Zuordnung von unbekannten Termen über String-Ähnlichkeiten funktioniert. Jedoch werden dabei viele falsche Verknüpfungen hergestellt, weil die Zuweisung ausschließlich auf Grundlage der Zeichenkette stattfindet. Eine andere Möglichkeit, die mehr semantisches Wissen in die Zuweisung einfließen lässt, besteht darin, Word Embeddings für die Kategorisierung von unbekannten Ausdrücken zu verwenden. Wie bereits unter 4.3. gezeigt kann auf Grundlage der Stellenanzeigen ein Vektorraummodell erstellt werden, das die semantischen Beziehungen zwischen den Wörtern in Stellenanzeigen abbildet. Dieses Word Embedding soll nun genutzt werden, um Synonyme zu identifizieren.

6.1. Aufgabenstellung und Software-Umsetzung

Unter 4.2. wurde bereits beschreiben, dass sowohl die ESCO als auch das AMS-Qualifikations-Barometer Konzept-zentriert organisiert sind und Synsets enthalten. Für diese Synsets soll nun überprüft werden, ob die enthaltenen Ausdrücke auch im Word Embedding nah beieinander liegen. Sollte dies zutreffen, so ist es umso wahrscheinlicher das unbekannte Kompetenzen, die sich im Word Embedding ebenfalls in der Nachbarschaft von bekannten Kompetenzen aufhalten, fehlerfrei in die Taxonomie eingeordnet werden können.

Um zu prüfen, inwiefern die Synonyme aus der Taxonomie und das aus den TextKernel-Stellenanzeigen entwickelte Word Embedding aufeinander abgestimmt sind, wird für alle Ausdrücke in einem Synset geprüft, ob die nächsten Nachbarn des Ausdrucks mit den anderen Ausdrücken im Synset übereinstimmen[19]. Dabei kann nur für die Ausdrücke eine Aussage getroffen werden, die sowohl in der Taxonomie als auch im Word Embedding enthalten sind. Damit die Überschneidung möglichst groß ist, werden die Ausdrücke des Word Embeddings und die der Taxonomie ungeachtet ihrer Groß- bzw. Kleinschreibung verglichen. Dementsprechend kann es auch geschehen, dass auf einen Ausdruck in der Taxonomie zwei Wortvektoren gleichzeitig zutreffen. Nun wird geprüft, bei welchen Synsets mehr als zwei Ausdrücke durch Wortvektoren repräsentiert werden, weil die Evaluation nur für diese Synonyme durchgeführt werden kann. Für diese Ausdrücke wird nun der Nachbarschaftsrang im Word Embedding berechnet. Der Rang zwischen den beiden Lexemen l_1 und l_2 wird dabei

[19] siehe classification/find_synonyms.py

sowohl mit dem Rang von l_2 in der Nachbarschaft von l_1 als auch mit dem Rang von l_1 in der Nachbarschaft von l_2 berechnet, da die Rangbeziehung nicht keine symmetrische Eigenschaft wie beispielsweise die Cosinus-Ähnlichkeit ist und die beiden Ränge deshalb nicht zwingend gleich sind.

6.3. Evaluation

Im Word Embedding sind insgesamt 3769 Wortvektoren, wenn man gleiche Wörter mit unterschiedlichen Groß- und Kleinschreibungen nicht zusammenfasst. Tut man dies, so ergeben sich 3436 unterschiedliche Ausdrücke. Aus den Synsets, die im AMS-Qualifikationsbarometer enthalten sind, entstehen 100 Ausdruckspaare, für die jeweils der Nachbarschaftsrang bestimmt werden kann. Die Nachbarschaft dieser Paare liegt durchschnittlich auf Rang 268 der nächsten Nachbarn und im Median beim 59. Rang. Deutlich ist, dass der überwiegende Teil der Synset-Ausdrücke über eine Nachbarschaft vom Rang 200 oder niedriger verfügt (siehe Abbildung 8). Das stärkt die Annahme, dass die in Synsets gebündelten Ausdrücke auch in den Stellenanzeigen enge paradigmatische Beziehungen haben.

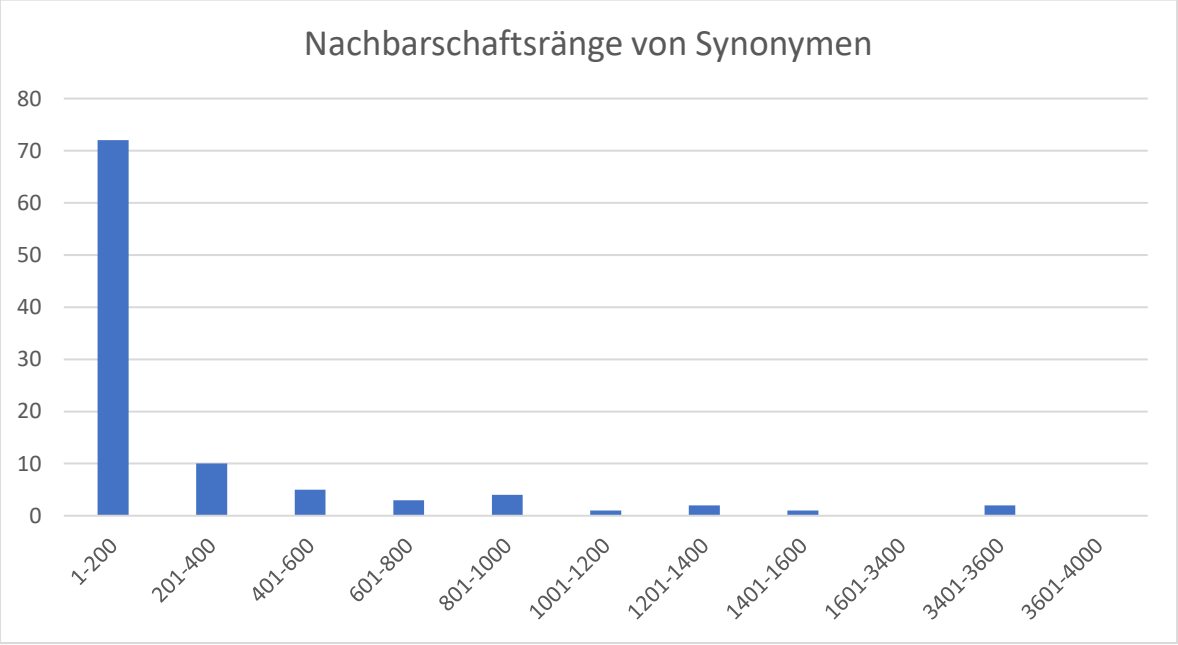

Abbildung 8: Häufigkeitsverteilung der Nachbarschaftsränge zwischen Synonymen

Eine sehr enge Nachbarschaft besteht beispielsweise bei *Dienstleistungsorientierung* und *Serviceorientierung*. Für beide Lexeme ist das jeweils andere Lexem der nächste Nachbar im Word Embedding. Dies trifft auch auf *Fortbildung* und *Weiterbildung* zu. Anders sieht es zum Beispiel bei *Büro* und *Sekretariat* aus, die jeweils der 1037. bzw. 3520. Nachbar des anderen sind. Beide Ausdrücke sind im AMS-Qualifikationsbarometer dem Konzept *Büroarbeitskenntnisse* zugeordnet. Insgesamt ist bei 28 Ausdruckspaaren der Rang höher als

200. Ein Auszug dieser Paare findet sich in Tabelle 3. Dabei fällt zunächst auf, dass hier oft die Wortarten der beiden Ausdrücke nicht übereinstimmen. Das spricht dafür, dass sie über keine paradigmatische Beziehung verfügen, obwohl sie im Qualifikations-Barometer als Synonyme geführt werden.

Tabelle 3: Beispiele für weit entfernte Synset-Lexeme

Lexem 1	Lexem 2	Rang
Einsatzbereitschaft	motiviert	205
Maschinenbau	Maschinen	264
Empathie	Verständnis	334
Freundlichkeit	herzlich	334
Verständnis	Einfühlungsvermögen	343
Verantwortungsbewusstsein	eigenverantwortlich	700
Büro	Sekretariat	845
Freundlichkeit	Herzlich	3520

Auch eine abweichende Groß- und Kleinschreibung kann Ursache für die große Distanz im Vektorraum sein, wie zum Beispiel bei *Herzlich* und *Freundlichkeit*. In der Variante, in der das Adjektiv kleingeschrieben wird, sinkt der Rang von 3520 auf 334. Außerdem könnte die Distanz zwischen *Empathie* und *Verständnis* ein Anzeichen dafür sein, dass *Verständnis* auch in anderen Zusammenhängen wie beispielsweise in *Verständnis für mathematische Zusammenhänge* auftreten kann. Dies deutet darauf hin, dass mehrdeutige Ausdrücke in diesem Ansatz ebenfalls eine gesonderte Rolle spielen.

7. Fazit & Ausblick

Diese Arbeit sollte zeigen, wie Wörter in Stellenanzeigen mit Bedeutung verknüpft werden können. Durch diese Verknüpfung wird es möglich, Stellenausschreibungen semantisch zu annotieren und so den Arbeitsmarkt besser zu analysieren. Im Rahmen der Untersuchung von Sprache in Stellenanzeigen wurde klar, dass im NLP-Bereich domänenspezifische Modelle sinnvoll wären. Diese können in Anschluss an diese Arbeit erstellt werden, beispielsweise für das Dependenz-Parsing oder das POS-Tagging. Die Verbreitung dieser Modelle würde vermutlich auf einige Resonanz stoßen, weil viele Projekte, die Arbeitsmarktanalysen betreiben bereits auf computerlinguistische Methoden zurückgreifen (siehe 3.2.). Für die Entwicklung dieser Modelle könnte das wesentlich größere Korpus von der Bundesagentur für Arbeit genutzt werden, jedoch wäre die Annotation der Trainingsdaten aufwendig. Außerdem könnte auf Grundlage des BA-Korpus ein Word Embedding wie unter 4.3. erzeugt werden, mit dem die anfänglichen Hypothesen über die Verschiebung der Wortvektoren im Vergleich zu Vektoren aus anderen Korpora überprüft werden können. Dieses Wort Embedding könnte auch genutzt werden, um Synonyme für die neu extrahierten Kompetenzen in den Taxonomien zu finden (siehe Kapitel 6). In Bezug darauf sollte erwähnt werden, dass diese Arbeit lediglich Kompetenzausdrücke bearbeitet hat, die nur ein Wort umfassen. Deshalb muss der Ansatz in Zukunft sowohl im Bereich der Disambiguierung als auch bei der Identifizierung von Synonymen für Mehrwortausdrücke erweitert werden. Das spielt vor allem bei der Berechnung des Word Embeddings eine Rolle, weil hier in der Vorverarbeitung neu definiert werden muss, wo Wortgrenzen liegen (siehe 2.4.). In Bezug auf die Disambiguierung könnte ferner untersucht werden, wie Informationen über die dependenziellen Partner der Kompetenzen aus Thesauri wie beispielsweise *GermaNet* dabei helfen können, die Disambiguierung zu verbessern (siehe Kapitel 5). Außerdem sollten beide vorgestellten Taxonomien dahingehend untersucht werden, ob ihre Bildung von Synonymen den hier vorgestellten Überlegungen dazu entsprechen, weil sich in Kapitel 6 zeigte, dass das Word Embedding der Stellenanzeigen und die Synsets der Kompetenz-Taxonomien nur teilweise übereinstimmen.

Literaturverzeichnis

Agirre, Eneko; Edmonds, Philip (2007): Word sense disambiguation. Algorithms and applications. Dordrecht: Springer (Text, speech and language technology, 33).

Agirre, Eneko; Martínez, David; Lacalle, Oier López de; Soroa, Aitor (2006): Two graph-based algorithms for state-of-the-art WSD. In: Dan Jurafsky und Eric Gaussier (Hg.): COLING-ACL 2006. EMNLP 2006, 2006 Conference on Empirical Methods in Natural Language Processing : 22 - 23 July 2006, Sydney, Australia ; proceedings of the conference. the 2006 Conference. Sydney, Australia, 7/22/2006 - 7/23/2006. Association for Computational Linguistics; Conference on Empirical Methods in Natural Language Processing; EMNLP 2006. Stroudsburg, Pa.: Association for Computational Linguistics (ACL), S. 585–593.

Calanca, Federica; Sayfullina, Luiza; Minkus, Lara; Wagner, Claudia; Malmi, Eric (2019): Responsible team players wanted: an analysis of soft skill requirements in job advertisements. In: *EPJ Data Sci.* 8 (1), S. 1–20. DOI: 10.1140/epjds/s13688-019-0190-z.

Clark, Kevin; Manning, Christopher D. (2016): Deep Reinforcement Learning for Mention-Ranking Coreference Models. In: Empirical Methods on Natural Language Processing. Online verfügbar unter https://nlp.stanford.edu/pubs/clark2016deep.pdf.

Cowie, Jim; Lehnert, Wendy (1996): Information extraction. In: *Commun. ACM* 39 (1), S. 80–91. DOI: 10.1145/234173.234209.

Cruse, D. Alan (2000): Aspects of the micro-structure of word meanings. In: Yael Ravin und Claudia Leacock (Hg.): Polysemy. Theoretical and computational approaches. Oxford: Oxford Univ. Press (/Oxford linguistics]), S. 30–51.

Dengler, Katharina; Matthes, Britta (2015): Folgen der Digitalisierung für die Arbeitswelt. Substituierbarkeitspotenziale von Berufen in Deutschland. Nürnberg (IAB-Forschungsbericht). Online verfügbar unter http://hdl.handle.net/10419/146097.

Fischer, Frank; Strötgen, Jannik (2017): Corpus of German-Language Fiction (txt). Online verfügbar unter https://figshare.com/articles/Corpus_of_German-Language_Fiction_txt_/4524680.

Geduldig, Alena (2017): Muster und Musterbildungsverfahren für domänenspezifische Informationsextraktion. Ein Bootstrapping-Ansatz zur Extraktion von Kompetenzen aus

Stellenanzeigen. Master Thesis. Universität zu Köln, Köln. Institut für Linguistik. Online verfügbar unter http://dh.uni-koeln.de/sites/spinfo/arbeiten/Masterthesis_Alena.pdf.

Glück, Helmut; Rödel, Michael (Hg.) (2016): Metzler Lexikon Sprache. 5., aktualisierte und berarbeitete Auflage. Stuttgart: J.B. Metzler. Online verfügbar unter https://ebookcentral.proquest.com/lib/gbv/detail.action?docID=4709275.

Han, Jiawei; Kamber, Micheline; Pei, Jian (2012): Data mining. Concepts and techniques. 3. ed. Amsterdam: Elsevier/Morgan Kaufmann (The Morgan Kaufmann series in data management systems). Online verfügbar unter https://ebookcentral.proquest.com/lib/subhh/detail.action?docID=729031.

Hermes, Jürgen; Schandock, Manuel (2016): Stellenanzeigenanalyse in der Qualifikationsentwicklungsforschung. Die Nutzung maschineller Lernverfahren zur Klassifikation von Textabschnitten. Bonn: Bundesinstitut für Berufsbildung.

Hirst, Graeme (1987): Semantic interpretation and the resolution of ambiguity. Cambridge: Cambridge University Press.

Iacobacci, Ignacio; Pilehvar Taher, Mohammad; Navigli, Roberto (2016): Embeddings for Word Sense Disambiguation: An Evaluation Study. In: Katrin Erk und Noah A. Smith (Hg.): Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany. Stroudsburg, PA, USA: Association for Computational Linguistics, S. 897–907.

Jones, Dale (2017): ESCO handbook. European skills, competences, qualifications and occupations. First edition. Luxembourg: Publication Office of the European Union.

Li, Keqian; Zha, Hanwen; Su, Yu; Yan, Xifeng (2018): Concept Mining via Embedding. In: 2018 IEEE International Conference on Data Mining (ICDM). 17-20 Nov. 2018. 2018 IEEE International Conference on Data Mining (ICDM). Singapore, 11/17/2018 - 11/20/2018. Los Alamitos, CA: IEEE Computer Society, Conference Publishing Services, S. 267–276.

Markowitsch, Jörg; Plaimauer, Claudia; Gaubitsch, Reinhold (2007): New developments in the early identification of skill needs in Austria: the AMS skills barometer. In: Olga Strietska-Ilina (Hg.): Systems, institutional frameworks and processes for early identification of skill needs. International conference on systems, institutional frameworks and processes for early

identification of skill needs. Luxembourg: Office for Official Publ. of the European Communities (Cedefop panorama series, 135), S. 191–200.

Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey (2013a): Efficient Estimation of Word Representations in Vector Space. Online verfügbar unter http://arxiv.org/pdf/1301.3781v3.

Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg; Dean, Jeffrey (2013b): Distributed Representations of Words and Phrases and Their Compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. USA: Curran Associates Inc (NIPS'13), S. 3111–3119. Online verfügbar unter http://dl.acm.org/citation.cfm?id=2999792.2999959.

Neuefeind, Claes (2019): Muster und Bedeutung: Bedeutungskonstitution als kontextuelle Aktivierung im Vektorraum: Modern Academic Publishing.

Prasad, K.N.S.S.V.; Saritha, S.K.; Saxena, Dixa (2017): A Survey Paper on Concept Mining in Text Documents. In: *IJCA* 166 (11), S. 7–10. DOI: 10.5120/ijca2017914143.

Resnik, Philip (1997): Selectional Preference and Sense Disambiguation. In: Tagging Text with Lexical Semantics: Why, What, and How?, S. 52–57. Online verfügbar unter https://www.aclweb.org/anthology/W97-0209.

Rösiger, Ina (2019): Computational modelling of coreference and bridging resolution. Unter Mitarbeit von Universität Stuttgart.

Rothe, Sascha; Schütze, Hinrich (2015): AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes.

Sahlgren, Magnus (2006): The word-space model. Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Zugl.: Stockholm, Univ., Diss., 2006. Stockholm: Dep. of Linguistics Stockholm Univ (SICS dissertation series, 44).

Schwarz, Monika; Chur, Jeannette (2007): Semantik. Ein Arbeitsbuch. 5., aktualisierte Aufl. Tübingen: Gunter Narr (Narr-Studienbücher).

Stohr, Daniel Christoph (2019): Die beruflichen Anforderungen der Digitalisierung hinsichtlich formaler, physischer und kompetenzspezifischer Aspekte. Eine Analyse von

Stellenanzeigen mittels Methoden des Text Minings und Machine Learnings (Sozialökonomische Schriften).

Turney, P. D.; Pantel, P. (2010): From Frequency to Meaning: Vector Space Models of Semantics. In: *jair* 37, S. 141–188. DOI: 10.1613/jair.2934.

Zhong, Zhi; Ng, Hwee Tou (2010): It Makes Sense: A Wide-coverage Word Sense Disambiguation System for Free Text. In: Proceedings of the ACL 2010 System Demonstrations. Stroudsburg, PA, USA: Association for Computational Linguistics (ACLDemos '10), S. 78–83. Online verfügbar unter http://dl.acm.org/citation.cfm?id=1858933.1858947.

Anhang

A Anleitung zur beigefügten Implementation

Dieser Arbeit ist – entweder als CD-ROM oder Zip-Datei – ein Archiv beigefügt, welches die in dieser Arbeit verwendeten Skripte sowie Teile der zur Ausführung benötigten Daten beinhaltet. Für alle Daten, die nicht im Archiv enthalten sind, gibt es entsprechende Hinweise, wo diese Daten, zum Beispiel die Word Embedding-Modelle, heruntergeladen werden können.

MA-ConceptMining

classification: Skripte zur Klassifikation und zur Identifizierung von Synonymen

data

corpus-of-german-language-fiction: steht unter folgendem Link zum Download bereit: https://figshare.com/articles/Corpus_of_German-Language_Fiction_txt_/4524680/1 ZIP-Datei entpacken und entsprechend einfügen

embeddings

german.model: Word Embedding, das aus Wikipedia-Texten und Zeitungsartikeln erstellt wurde. Steht unter folgendem Link zum Download bereit: https://devmount.github.io/GermanWordEmbeddings/ (unter Download/Model/german.model)

german_jobad.model: Word Embedding, das aus Stellenanzeigen gebildet wurde

german_inter.model & german_jobad_inter.model: Modelle der Wortvektoren, die eine Schnittmenge zwischen german_jobad.model und german.model bilden

jobads

classified_sentences.db: SQLite-Datei mit Sätzen und extrahierten Kompetenzen (teilweise klassifiziert und entsprechend mit 0 oder 1 annotiert)

text_kernel_jobads.db: vollständige Stellenanzeigen

skills

AMS_CategorizedCompetences.db: SQLite-Datei mit Auszug des AMS-Qualifikationsbarometers

esco_skills_de.csv: CSV-Datei mit Kompetenz-Taxonomie der ESCO

exploration: Skripte zur linguistischen Exploration

inout: Skripte mit Hilfsfunktionen für Input und Output

preprocessing: Skripte mit Hilfsfunktionen für die Vorverarbeitung und Skript zur Bildung der Word Embedding-Schnittmenge (Kapitel 4.3.)

Die Skripte wurden in der Entwicklungsumgebung PyCharm mit Python 3.8 geschrieben. Für die Ausführung sind folgende externe Bibliotheken notwendig:

- gensim 3.8.1
- spacy 2.2.3
- scikit-learn 0.22.1
- nltk 3.4.5

B Eidesstattliche Versicherung

Hiermit versichere ich an Eidesstatt, dass ich die vorliegende Arbeit selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß ausveröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche unter Angabe der Quelle kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oderauszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Hürth, den 24.03.2020 Unterschrift: