UNIVERSITÄT LEIPZIG

Fakultät für Mathematik und Informatik

Institut für Informatik

Diplomarbeit

Thema: Vererbungsalgorithmen von semantischen Eigenschaften auf Assoziationsgraphen und deren Nutzung zur Klassifikation von natürlichsprachlichen Daten

vorgelegt von

Stefan Bordag

geb. am: 03. Juni 1978

Studiengang Informatik

Leipzig, Juni 2002

Zusammenfassung

Ziel dieser Arbeit ist es, Kollokationen auf Satzbasis aus dem Wortschatz-Lexikon Leipzig als Datenbasis nehmend, ein Verfahren zu entwickeln, welches die den Satzkollokationen immanenten Beziehungen zwischen den Wörtern erkennen und handhaben kann. Weiterhin ist es Ziel, diese Beziehungen für wortbedeutungsorientierte Klassifikationsverfahren zu erschließen und deren unmittelbare Anwendung zu demonstrieren, indem Sachgebietszuweisungen über diese Beziehungen weitervererbt werden können.

Es wird gezeigt, dass Cluster in den Satzkollokationen mit einer Approximation für die maximale Clustersuche mit rechnerisch geringem Aufwand gefunden werden können, wenn diese als ein Graph mit der seit kurzem untersuchten small-world Eigenschaft betrachtet werden. Es wird daraufhin ein Disambiguierungsverfahren konstruiert, welches Beziehungen zwischen einem Wort und seinen unmittelbar angrenzenden Clustern berechnet, wobei die verschiedenen Cluster den diversen Gebrauchskontexten und damit auch unter anderem den Bedeutungen des Wortes entsprechen.

Dieses Disambiguierungsverfahren dient dann als Grundlage für den Entwurf eines Sachgebietsklassifizierungsverfahrens, welches zu einer inhaltlich homogenen Wortgruppe, zum Beispiel einem Sachgebiet, weitere passende Wörter finden kann.

Die vorgeschlagenen Verfahren wurden prototypisch implementiert und Beispiele werden auch im Hinblick auf eine Praxisanwendung diskutiert.

Inhaltsverzeichnis

1. Einleitung

Mit der beständig wachsenden Nutzung des Computers durch den Menschen und seiner nach dem Mooreschen Gesetz wachsenden reinen Rechenkapazität eröffnen sich stets neue Möglichkeiten und Anforderungen an die Kommunikation mit dem Computer. Es geht dabei nicht mehr nur darum, direkt einem Computersystem mitzuteilen, welche Aufgabe es durchführen soll. Hinzugekommen sind viele andere Anwendungen, wie etwa die automatische Analyse und Produktion natürlicher Sprache in geschriebener und gesprochener Form, automatische Übersetzung von einer natürlichen Sprache in eine andere, um das Finden von passenden Texten oder Textstücken aus einer großen Sammlung auf eine ebenfalls in natürlicher Sprache verfasste Anfrage hin und nicht zuletzt die direkte Interaktion des Menschen mit dem Computer.

Statistische Verfahren spielen dabei an allen solchen Stellen eine wachsende Rolle, an denen die Verarbeitung von großen Mengen von Wissen bzw. Daten benötigt werden, bei welchen Handarbeit zu viel Aufwand bedeuten würde und eine hundertprozentige Genauigkeit, welche mit statistischen Verfahren meist nicht erreicht werden kann, auch nicht benötigt wird. Das prominenteste Beispiel ist wohl die gewöhnliche Internetsuchmaschine, welche natürlichsprachliche Eingaben akzeptiert und daraufhin möglichst inhaltlich passende Dokumente findet.

1.1. Motivation

Die Motivation für die Arbeit entstand aus der Beobachtung, dass sich in einem großen Korpus von natürlichsprachlichen Daten makroskopische Effekte beobachten lassen, die in kleinen Beispielen nicht beobachtbar sind und die weiterführen als reine Auftrittswahrscheinlichkeiten, Trigrammanalysen oder andere bekannte Methoden. Es bestand daraufhin die Vermutung, dass diese Effekte sich nach genaueren Untersuchungen nutzbar machen lassen müssten und neue

Perspektiven in der automatischen Sprachverarbeitung eröffnen könnten.

Vor allem geht es um die Beobachtung, dass einige Wörter offensichtlich auch über längere Beobachtungsabschnitte hinweg statistisch signifikant oft gemeinsam in Sätzen auftreten. Daraus lassen sich Beziehungen zwischen den entsprechenden Wörtern modellieren, deren Grundlage allem Anschein nach die Bedeutung der einzelnen Wörter darstellt.

Die Beziehungen zwischen den Wörtern wiederum können als Grundlage genutzt werden, um manuell eingegebenes Wissen über einzelne Wörter auf neue Wörter vererben zu können, bzw. die Wörter nach diesem Schema zu klassifizieren.

1.2. Ziel der Arbeit

Das Ziel dieser Arbeit ist demnach:

1. Die Ausarbeitung eines Verfahrens, welches die Struktur der Beziehungen zwischen den Wörtern korrekt erkennen und handhaben kann.
2. Vorstellen von Vorschlägen und Beispielen von Klassifizierungsmöglichkeiten, die diese Struktur ausnutzen, sowie deren prototypische Implementierung.

Dabei werden Satzkollokationen als Grundlage genommen, um Beziehungen zwischen Wörtern zu modellieren und es wird Wissen über Sachgebietszugehörigkeiten von einzelnen Wörtern genutzt, um mit Hilfe dieses Wissens Wörter zu klassifizieren, die noch keinem Sachgebiet zugewiesen sind. Der Schwerpunkt der Arbeit liegt dabei bei dem ersten Punkt, unter anderem auch deswegen, weil Sachgebiete nur eine der vielen möglichen Daten sind, die über neue Wörter generalisiert werden können, bzw. nach denen die Wörter klassifiziert werden.

Es kann im Umfang dieser Arbeit zwar keine empirische Auswertung aller sich öffnenden Möglichkeiten durchgeführt werden,

doch werden sie an den entsprechenden Stellen angesprochen und exemplarisch vorgestellt.

1.3. Gliederung

Im Verlauf dieser Arbeit wird der Leser zunächst über die vorhandenen Daten informiert, wonach die zu findende Struktur definiert wird und gezeigt wird, dass sie in den gegebenen Daten existiert. Danach wird ein Verfahren entwickelt, diese Struktur auszunutzen, woraus eine Anwendung folgt. Diese Anwendung wird dann für die Klassifizierungen genutzt.

Zunächst wird im folgenden Kapitel genauer auf die Problematik der Sachgebiete mit allen relevanten Folgen eingegangen, um Möglichkeiten, aber auch Restriktionen zu diskutieren. Im dritten Kapitel werden dann die Daten mit einigen für diese Arbeit relevanten Hintergründen beleuchtet, welche als Grundlage für diese Arbeit dienen. Es wird dabei vor allem auch auf die Satzkollokationen eingegangen.

Das vierte Kapitel setzt sich mit der den Satzkollokationen inhärenten Struktur auseinander und liefert den theoretischen Hintergrund, während das fünfte Kapitel einen Weg zeigt, diese Struktur handhaben zu können.

Das sechste Kapitel behandelt eine unmittelbare Anwendung des vorher definierten Verfahrens und diese Anwendung wird dann direkt im letzten Kapitel für die Klassifizierungen genutzt.

2. Semantisches Wissen

Für eine automatische Klassifikation von Wörtern nach einem beliebigen, möglichst günstigen semantischen Kriterium bot sich die in dem deutschen Wortschatz-Lexikon vorhandene Sachgebiets-Datenbank an, da in ihr für eine kleine Teilmenge der gesamt vorkommenden Wörter eine Einteilung in Sachgebiete vorhanden war. Es ist klar, dass eine Einteilung einer Menge von Sprachmaterial, namentlich hier von Vollformenwörtern, nach semantischen Gesichtspunkten durch ein Programm verlangt, dass diesem Programm Regeln geliefert werden, nach denen es diese semantischen Gesichtspunkte ausnutzen kann, um für neue Wörter zu entscheiden, zu welcher Klasse sie gehören sollen. Im Weiteren wird jede Vollform eines Wortes als einzelnes und unterschiedliches Wort bezeichnet. Mit Klassifikation ist gemeint, dass eine Menge von Wörtern oder all-gemeiner Objekten nach einem zu spezifizierenden Schema eingeteilt wird, wobei ein Wort auch mehreren Kategorien zugeteilt werden kann.

Herkömmliche, regel- und musterbasierte Ansätze, wie sie in der Automatischen Sprachverarbeitung für spezielle Aufgaben häufig eingesetzt werden, um zum Beispiel automatisch männliche von weiblichen Vornamen zu unterscheiden, müssen hier versagen, da sie den Fakt ausnutzen, dass im Deutschen und den umliegenden Sprachen eine Unterscheidung zwischen männlich und weiblich durch die Derivation eingeführt wird, wie zum Beispiel ‘Heiko – Heike’ oder ‘Mario – Maria’ oder eine solche Unterscheidung durch Mustererkennung möglich ist, wie zum Beispiel, dass alles, was mit ‘Alf-’, ‘Diet-’ oder ‘Hans-’ beginnt, mit größter Wahrscheinlichkeit männlich sein wird.

Die Derivation, sowie die Flexion, die eine ähnlich für die vorliegende Aufgabe eingeschränkte Rolle spielt, sind allerdings nur sehr begrenzt an die Semantik gebunden. Sie können den Modus eines Wortes verändern (aus ‘kann’ wird ‘könnte’), sie können die Bedeutung eines Wortes gar umkehren (‘un-’ Präfix), aus einem Substantiv ein

Verb derivieren (aus ‘Wasser’ wird ‘wässern’), oder eine Reihe anderer Einflüsse auf die Bedeutung eines Wortes ausüben, doch die eigentliche Einbettung in das Bedeutungsgefüge kann durch die reine Flexion oder Derivation nicht herausgefunden werden, da es sich immer um Regeln handelt, die auf einer Klasse von Wörtern funktionieren und vor allem bei der Flexion unabhängig von der lexikalischen Wortbedeutung sind. Damit ist auch determiniert, welche Klasse von Aufgaben durch regelbasierte Ansätze gelöst werden können und welche nicht. Es können mit derartigen formbasierten Verfahren Unterscheidungen zwischen Klassen von Wörtern entschieden werden – aber es können nicht wortbedeutungsspezifische Entscheidungen getroffen werden.

Dadurch wird klar, dass für eine echte bedeutungsorientierte Einteilung in einer beliebigen Form auch echtes Wissen über Bedeutung benötigt wird – das heißt aber nicht, dass das Programm „verstehen“ muss, worum es geht. Dies schien mit der Sachgebietsdatenbank zusammen mit der Kollokationsdatenbank im Wortschatz-Lexikon gegeben zu sein, da mit der Einteilung einer Menge von Wörtern zu einem Sachgebiet die Aussage getroffen wurde, dass diese Wörter eine Bedeutungsgemeinsamkeit besitzen, doch führte die Sachgebietsdatenbank zunächst mehr Probleme ein, als sie löste und darauf soll in diesem Kapitel genauer eingegangen werden. Im nächsten Kapitel wird dann auf die restlichen Daten eingegangen, vor allem aber auf die Kollokationen. Diese entsprechen dem strukturalistischen Modell dahingehend, als dass sich der Inhalt einer Einheit einzig aus dem Gebrauch, bzw. den Beziehungen zu anderen Einheiten berechnen lässt.

2.1. Sachgebietsklassifikation

In nahezu jeder größeren Sammlung von Information, bzw. Wissen, oder auch allgemeiner von Objekten wird von der sammelnden Institution (z. Bsp. einem Menschen) eine Ordnung eingeführt, die sich nicht lediglich nach alphabetischen Listen von Autoren – oder Titel-

namen (bei Bibliotheken) also nach der Form der zu sammelnden Objekte richtet. Vielmehr richtet sich diese Ordnung nach dem Wert, Bedeutung, Funktion oder Extension der gesammelten Objekte. Handelt es sich Beispielsweise um Bücher, wird nach Themengebieten eingeteilt, welche wiederum in viele verschiedene nahezu beliebig fein verästelte Kategorien zergliedert werden, je nach Größe der Sammlung. Handelt es sich etwa um einen Supermarkt, wird nach Verbrauchsfunktion eingeteilt und dann sind mit einem mal alle Bücher in der gleichen Kategorie. Handelt es sich dagegen um ein Information Retrieval System, welches Dokumente aufgrund von Ähnlichkeit klassifiziert und damit ebenfalls versucht, eine bedeutungsorientierte Ordnung über der betrachteten Dokumentmenge zu definieren, so wird auch dieses im Grunde eine „korrekte“ Sachgebietseinteilung sein, obwohl vielleicht nicht immer gesagt werden kann, was konkret dieses spezielle Sachgebiet bedeutet, aber das ist auch nicht immer von Relevanz[i].

Da die Ähnlichkeitsfunktion in IR Systemen meist kaum mehr als ein gewichtetes Skalarprodukt über eine ausgewählte Menge von Termen (meistens Wörtern) und ihrem entsprechenden Vorkommen oder Nichtvorkommen im Dokument ist, handelt es sich auch hier lediglich um eine Approximation – das „Verstehen“ des Dokumentes wird simuliert und das „Wissen“ des Programms besteht eigentlich nur aus dem Vergleich des betrachteten Dokumentes mit anderen Dokumenten. Doch auch dieses höchst eingeschränkte Wissen scheint zumindest für einige spezielle Gebiete ausreichend zu sein, wie die Erfolgsraten derartiger Systeme zeigen und das liegt daran, dass einzelne Wörter und Phrasen in solchen Dokumenten signifikant häufiger vorkommen, die inhaltlich mit der ent-sprechenden Wortsemantik eng verknüpft sind, als es der Fall für die Gesamtmenge von Dokumenten ist.

---

[i] Siehe zum Beispiel bei Deerwester, Dumais und Harshman [DDH90], es ist dann die Rede von abstrakten Einteilungsvektoren.

Die bekannteste Sorte von Information Retrieval Systemen dieser Art (also statistische Klassifikation) sind wohl wie bereits erwähnt die Suchmaschinen im Internet (www.google.de, www.altavista.de, ...).

2.2. Probleme der Sachgebietsklassifikation

Bei sämtlichen Anwendungen von Kategorisierung zeigt sich deutlich vor allem die Uneinigkeit darüber, was eigentlich eine gute Einteilung, bzw. Kategorisierung sein soll. Offensichtlich sind die meisten Kategorien in der Sprache, bzw. unserem Weltwissen mindestens in einer hierarchischen Struktur, es gibt Oberbegriffe und Unterbegriffe und zu einem Oberbegriff kann man meist leicht wieder einen Oberbegriff finden. Insgesamt lassen sich mehrere verschiedene Beziehungen zwischen Wörtern beobachten, wie zum Beispiel 'is-a', 'part-of' oder andere.

Diese Beziehungen würden in eine hierarchische Struktur passen, wenn nicht die Möglichkeit bestünde, dass der Mensch dem gleichen Weltwissen problemlos zwei grundverschiedene hierarchische Ordnungen aufdefinieren kann und damit beweist, dass es für ihn in der Welt komplexere Assoziationen gibt, als solche, welche mit einem Baumgraphen beschrieben werden können. Nach näherer Betrachtung stellt sich demnach heraus, dass auch eine hierarchische (Baum-) Struktur nicht ausreichend ist, um die Zusammenhänge ausreichend beschreiben zu können.

Darüber hinaus entsteht noch ein weiteres Problem, und zwar überhaupt zu entscheiden, was genau eine ideale Sachgebietseinteilung ist. Soll ein Wort nun zu

- dem möglichst passenden Sachgebiet zugeteilt werden,
- möglichst vielen passenden Sachgebieten zugeteilt werden oder
- soll kontextabhängig entschieden werden, welche der beiden Varianten angewendet wird

Dabei spielt auch eine Rolle, welche Beziehungen zwischen den betrachteten Wörtern bestehen. Da es sich in diesem Fall eben um Wör-

ter handelt, muss beachtet werden, dass es zwischen ihnen paradigmatische Beziehungen gibt, wie etwa Synonymie, Opposition, Hyperonymie und Hyponymie. Ferner kann jedes einzelne Wort auch noch mehrere grundverschiedene Bedeutungen besitzen – also ambig sein (Homonymie).

Es ist deutlich, dass es weder Aufgabe, noch Ziel dieser Arbeit sein kann, eine perfekte Klassifizierung zu finden. Es ist eher erkennbar, dass die Bedürfnisse der Benutzer eines Klassifizierers sich wandeln können und dass angeraten ist, ein System oder einen Algorithmus zur Verfügung zu stellen, der sich möglichst gut an diese Bedürfnisse anpasst, wie bei Heyer & Haugeneder [Heyer95] beschrieben. Dies kann mit den in dieser Arbeit entwickelten Verfahren auch erreicht werden, siehe Kapitel 7.2. Es sollte also vielmehr ein gut dokumentiertes Modul darstellen, bei welchem bekannt ist, welche Art von Daten es benötigt und welche Art von Daten es liefert und dass die Ergebnisse dieses Moduls verlässlich in beliebigen weiteren Sprachprodukten weiterverwendet werden können, die an der gegebenen Funktionalität interessiert sind.

2.3. Sachgebiete im Wortschatz-Lexikon

Die Sachgebietsdatenbank, die bereits kurz erwähnt wurde, ist eine größere Sammlung von Klassifizierungen aus verschiedenen Quellen. Es wurden Vollformen von Wörtern jeweils einem oder mehreren Sachgebieten zugeordnet, wobei die Benennung der Sachgebiete wiederum aus einem oder mehreren bekannten Wörtern besteht. Aufgrund der verschiedenen Quellen existiert in dieser Datenbank eine starke Unausgewogenheit in der Klassifizierung. Während einige Gebiete sehr fein verästelt sind und sehr fachspezifische Begriffe besitzen, gibt es andere, die dagegen sehr grob sind und eher allgemeine Begriffe beinhalten.

Es stellt sich an dieser Stelle auch die Frage, welche Wörter eigentlich etwa zum Sachgebiet ‘Versicherung’ gehören sollen. In der Datenbank findet man unter anderem : ‘Haftpflichtversicherung, Ver-

sicherungsvertreter, Lebensversicherung, Unfallversicherung, Versicherungsmakler, Versicherungswirtschaft, Rückversicherung, Krankenversicherung, Pensionskasse, Versicherungskaufmann, ...'. Es hätten allerdings auch folgende nicht in dieser Datenbank aufgeführte sein können: 'Bemessungsgrenze, Ersatzkassen, Kostenerstattung, Sachschaden, Unfall, Deckungssumme, Schadensgutachten, ...'. Erstere sind eher Unterarten von Versicherungen, während zweitere eher in diesem Bereich oft gebrauchte Wörter darstellen.

Ist ein System verlangt, das alle diese Zuweisungen finden kann, unabhängig des zugrundeliegenden Prinzips, müsste in diesem System modular ein Verfahren für jedes einzelne verschiedene Prinzip existieren, nach welchem Wörter klassifiziert werden. In der vorliegenden Arbeit wird eines dieser Verfahren entwickelt und zwar eines, welches Wörter nach ihrem Gebrauchskontext zuweisen kann. Es wird davon ausgegangen, dass ein Sachgebiet durch die Gesamtheit der ihm zugewiesenen Wörter definiert ist. Das heißt, dass dieses Verfahren für Definitionen, wie sie weiter oben für 'Versicherung' gegeben waren, keine optimalen Ergebnisse liefern werden kann, weil die definierenden Wörter oft zu spezifisch und selten sind und damit keine statistischen Messungen für diese Wörter möglich sind.

3. Wortschatz-Lexikon

Die zweite wichtige Komponente der Ausgangsdaten für diese Arbeit stellen neben den Sachgebieten die Kollokationsdaten dar, auf die im Folgenden näher eingegangen wird.

Das Wortschatz-Lexikon in Leipzig ist aus einer Vielzahl von Quellen semi-automatisch aufgebaut. Das heißt, dass einige größere Zeitungen Deutschlands ('Die Zeit', 'TAZ', 'Der Spiegel', ...)[i] täglich eingespeist werden, dabei zunächst in ihre Sätze, danach in ihre Wörter zerlegt werden und anschließend beides gespeichert wird. Dabei werden die Wörter in der Form, in der sie ankommen, gespeichert, denn es handelt sich um ein Vollformenlexikon.

Weiterhin wird zu jedem Wort gezählt, wie oft es auftrat. Die Sätze werden ebenfalls komplett gespeichert, weil sie für diverse spätere Auswertungen benutzt werden können (die vorliegende Arbeit ist eine solche Analyse) und weil es nicht gegen Urheberrechte verstößt. Die Texte selbst dürfen leider nicht gespeichert werden und können somit nicht genutzt werden. Weiterhin wurden alle (urheberrechtlich und elektronisch) verfügbaren Fachlexika, Fachzeitschriften und Monographien aus unterschiedlichen Wissensgebieten (u. a. Medizin, Rechtswissenschaft, Informatik) eingespeist.

Dieser inzwischen gigantische Datenbestand wurde über Jahre hinweg (seit 1998) mit verschiedenen semi-automatischen Verfahren gepflegt, darunter fallen zählen z. Bsp.:

- Entfernung von ganzen Datenbeständen, wenn erkannt wurde, dass eine bestimmte Quelle im Schnitt sehr schlechte Daten liefert, also triviale Rechtschreibfehler, Dialektsprache, häufige Anglizismen, usw.

[i] Eine genaue Auflistung der Partner findet sich unter: http://wortschatz.uni-leipzig.de/html/partner.html

- Per Hand Einfügen von Mehrwortbegriffen, weil es nach wie vor keinen hinreichend guten Algorithmus gibt, der solche automatisch erkennen würde.
- Zuweisungen von natürlichsprachlichen Wortbeschreibungen.

3.1. Das Zipfsche Gesetz

Die frühesten statistischen und mit die einflussreichsten Untersuchungen großer Mengen natürlichsprachlicher Daten stammen von George K. Zipf und resultierten unter anderem in den drei Zipfschen Gesetzen:

Erstes Zipfsches Gesetz bezieht sich auf das Verhältnis zwischen dem Rang r und der Frequenz f eines Wortes, wobei der Rang eines Wortes die Platznummer des Wortes in einer nach Häufigkeit sortierten Wortliste darstellt[i] und die Frequenz die absolute Häufigkeit. Die Multiplikation dieser beiden Werte für ein beliebiges Wort ergibt dann eine Konstante k und demnach lautet dieses Gesetz:

$$r \cdot f \approx k$$

Dieses erste Zipfsche Gesetz hatte allerdings einen Vorgänger – den Französischen Mathematiker J. B. Estoup, welcher bereits vor Zipf für die Stenographie den gleichen Zusammenhang zwischen der Frequenz und dem Rang eines Wortes formulierte[ii] und wird daher manchmal auch Estoup-Zipfsches Gesetz genannt.

Spätere Untersuchungen, wie etwa Těšitelová & co. [TK87], zeigten außerdem, dass dieses Gesetz für den mittleren Bereich der nach Frequenz sortierten Wortliste zwar stimmt, aber sowohl für Wörter mit extrem hoher Frequenz, als auch für solche mit sehr niedriger Frequenz nicht mehr, weil vor allem bei letzteren zu beobachten ist, dass immer mehr Wörter mit der Frequenz 1 zu beobachten sind, aber ihr

---

[i] wobei Wörter mit gleicher Frequenz den gleichen Rang haben, im Gegensatz zu einem Platz in einer echten nach Frequenz sortierten Liste

[ii] siehe auch J. Černý [Černý96] S. 254

Rang immer mehr steigt, womit der berechnete Wert sich immer mehr von der Konstante entfernt.

Aber dies kann auch daran liegen, dass im oberen Frequenzbereich nicht genügend Wörter existieren, um die Werte zuverlässig werden zu lassen, und im unteren Frequenzbereich nicht genügend Sprachdaten beobachtet wurden.

Zweites Zipfsches Gesetz bezieht sich auf die Beziehung zwischen der Frequenz eines Wortes f und der Anzahl b der Wörter, die die gleiche Frequenz besitzen. Die Aussage an dieser Stelle ist, dass eine Multiplikation dieser beiden Werte für ein beliebiges Wort wieder eine Konstante k ergibt:

$$f \cdot b \approx k$$

Drittes Zipfsches Gesetz bezieht sich auf den Vergleich der Frequenz f eines Wortes mit der Anzahl seiner Bedeutungen m:

$$\frac{m}{\sqrt{f}} \approx k$$

Was auch so vereinfachend interpretiert werden kann, dass je häufiger ein Wort vorkommt, umso mehr Bedeutungen hat es, bzw. in umso mehr Kontexten ist es ein normal benutztes Wort. Dieses letzte Gesetz ist für diese Arbeit von großer Relevanz, denn wie später zu sehen sein wird, haben Wörter mit großer Frequenz in der Tat die Tendenz, in jedem Kontext mitzuwirken und werden daher getrennt behandelt oder einfach weggefiltert.

Diese Zipfschen Gesetze lassen sich mit leichter Wertevariation für die meisten bekannten Sprachen beobachten. Weitere Beobachtungen ergeben zum Beispiel, dass bei den sehr häufigen Wörtern eher Funktionswörter mit kaum eigenem Inhalt, Verben und Zählwörter auftreten, während je weiter man sich den selteneren Wörtern nähert, umso mehr trifft man Substantive, Komposita und Wörter mit sehr speziellen Bedeutungen und immer weniger Verben und Adjektive und diese sind dann auch eher fachspezifisch.

Folgende Abbildung wurde erstellt, indem aus dem Wortschatz-Lexikon, dessen Worteinträge eine von der Frequenz abhängige Wortnummer besitzen, in exponentiell wachsenden Abständen[i] die Wortartenverteilung gemessen wurde. Es wurden 4 grobe Kategorien gewählt: Substantive, Adjektive, Verben und Funktionswörter, wobei zu den letzteren die Präpositionen, Artikel, Konjunktionen, Pronomina, Pronominaladverbien, Partikel und spez. Verbformen gezählt wurden.

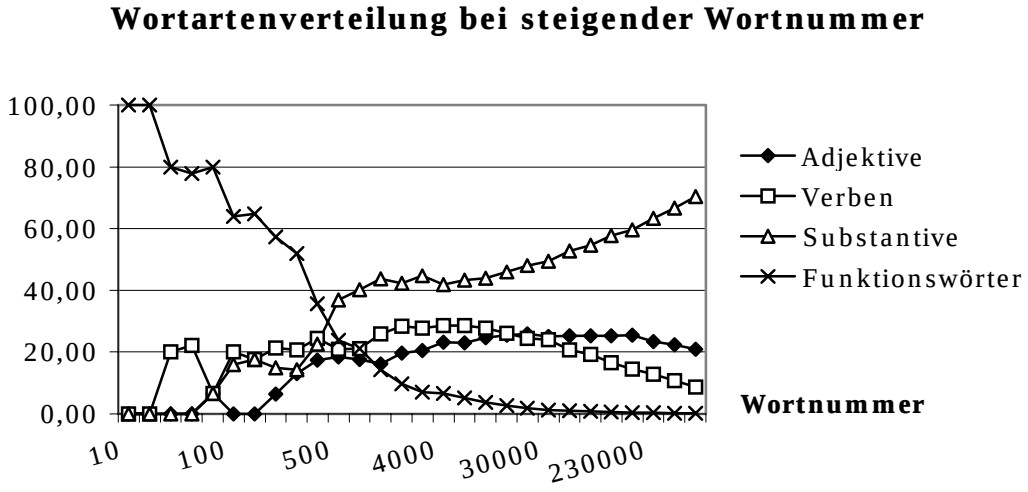


Abbildung 1

3.2. Bisherige Analysen

Weiterhin haben sich Mitarbeiter und Studenten der Abteilung für automatische Sprachverarbeitung um Auswertungsmöglichkeiten bemüht, die darin resultierten, dass nunmehr unter anderem online (http://wortschatz.uni-leipzig.de) eingesehen werden kann:

1. Welche Wörter statistisch signifikant oft linke Nachbarn von welchen anderen sind.

[i] Nach der Vorschrift $5 \cdot 1{,}2^i$, wobei die ersten drei Datenpunkte zusammengelegt wurden.

2. Welche Wörter statistisch signifikant oft rechte Nachbarn von welchen anderen sind.
3. Welche Wörter überhaupt statistisch signifikant zusammen im gleichen Satz auftreten.

Solche Auftrittsanalysen zeigen mindestens, dass die zwei gefunden Wörter „etwas“ mit einander zu tun haben, obwohl leider nicht gesagt werden kann, „was“ genau. Man nennt diese zwei Wörter eine Kollokation, wobei man Unterscheidungen zwischen Nachbarschaftskollokationen und Satzkollokationen trifft, wobei sich die Nachbarschaftskollokationen naturgemäß noch in linke und rechte Nachbarn aufteilen.

In der Diplomarbeit von F. Schmidt [Schmidt99], sowie in der Dissertation von A. Lehr [Lehr96] findet sich jeweils ein guter Überblick über Kollokationen. Allgemein ist dieser Begriff linguistisch vorbelegt und Kookkurenz wäre ein besserer Begriff, allerdings hat sich „Kollokation“ in Leipzig eingebürgert und wird daher auch in dieser Arbeit in dieser Form benutzt.

Bei den Satzkollokationen lassen sich zu einem Wort von Synonymen über Antonyme bis hin zu Hyperonymen und Hyponymen mehr oder weniger alle Arten von semantischen Beziehungen entdecken. Weiterhin treten einfach aktuelle Wörter wie z. Bsp. ‘Kohl’ signifikant häufig in gleichen Sätzen wie ‘Sitzung’, ‘Politik’ u. ä. auf (Homographie).

Es wird deutlich, dass die unterschiedlichen Kollokationsarten auch prinzipiell unterschiedliche Ergebnismengen haben müssten. Während bei den linken Nachbarschaftskollokationen eher grammatisch passende Wörter auftreten, also vor Substantiven eher Verben, Adjektive und Artikel, sind es vor Verben wiederum bevorzugt Adverbien, Adjektive oder Substantive. Bei den Satzkollokationen handelt es sich eher um semantisch passendere Wörter.

Das liegt auch daran, dass im Deutschen im Vergleich zu anderen, flektiveren Sprachen, die Wortfolge stabil ist, d.h. nicht alle Umord-

nungen von Wörtern in einem Satz ergeben einen korrekten Satz und wenn doch, so haben sie eine jeweils andere Bedeutung.

Die Eigenschaft der Satzkollokationen ist besonders für diese Arbeit von Relevanz, weil sie hier zur Auswertung herangezogen werden. Allgemein kann man die Satzkollokationen auch als einen Graphen betrachten, bei dem die Knoten die Wörter sind und die Kanten, zwei Wörter verbindend, die Kollokationen. Dieser Graph ist bei den Satzkollokationen ungerichtet, aber die Kanten haben Gewichte – die Kollokationssignifikanz.

3.3. Die Quellen und ihr Inhalt

„Was man sät, das erntet man auch“ – ein altes Sprichwort und trifft auch hier zu - beim Wortschatz-Lexikon der Deutschen Sprache. Wie bereits erwähnt wurde, besteht der größte Teil der Daten aus Tageszeitungen. In den Zeitungen gibt es allerdings nur eine sehr eingeschränkte Auswahl von Themen, die dafür immer wiederkehren und notgedrungen natürlich immer wieder das gleiche Sprachmaterial benutzen müssen. Ein gutes Beispiel ist der Sportteil einer Zeitung. Er nimmt nahezu in jeder Zeitung einen nicht zu unterschätzenden Teil jeder Ausgabe ein und davon wieder ist ein großer Teil dem Fußball gewidmet. Damit tritt automatisch eine Verzerrung ein, denn das Wort Spiel befindet sich nun plötzlich fast ausschließlich im Fußballkontext, während alle anderen Bedeutungen oder möglichen Kontexte des Wortes Spiel, wie zum Beispiel das Bühnenspiel, stark unterrepräsentiert sind.

Sollte nun eine semantische Analyse anhand der Kollokationen durchgeführt werden, dann sollte es natürlich auch nicht verwundern, dass für eine gewisse Menge von Wörtern der Kontext im Wortschatz-Lexikon ein etwas anderer ist, als die Intuition erwarten ließe. Das liegt daran, dass sich die „Welt“ des Wortschatz-Lexikons nahezu einzig aus Zeitungen zusammensetzt, während „unsere“ Welt noch viel mehr Facetten und Informationsquellen besitzt. Auch ist nicht ganz klar, was ein Mensch wohl unter ‘Spiel’ verstehen würde,

wenn seine einzige Informationsquelle tag-täglich aus einigen Zeitungen bestünde. In der Kollokationsmenge eines Wortes können sich demnach Wörter befinden, die unter anderem aufgrund von folgenden Faktoren im gleichen Satz auftreten:

Zwischen den Wörtern existiert eine

- Oppositionsbeziehung
- Hyperonymie- / Hyponymiebeziehung
- Grammatische Bedingtheit (‘einteilen’ – ‘teilte … ein’)

Wörter treten auch im gleichen Satz auf, weil

- das aktuelle Geschehen in den Zeitungen dazu führt, dass zwei eigentlich unabhängige Wörter plötzlich in einem Kontext genannt werden (‘Rinder’ – ‘Wahnsinn’)
- sie zusammen eine Redewendung ergeben

Synonyme hingegen werden äußerst selten im gleichen Satz auftreten.

Doch das eigentliche Problem liegt nicht darin, dass alle diese Faktoren wirken, sondern darin, dass sie alle scheinbar ohne erkenntlichem System gemischt in der Kollokationsmenge auftreten und die Aufgabe wäre nun, Dichotomien zu finden, aufgrund derer diese Menge in sinnvolle Teilmengen zerlegt werden könnten, was das Anliegen dieser Arbeit darstellt.

Unabhängig davon, ob es gelingt, diese Mengen nach allen oder zumindest einigen Kriterien zu trennen, sollte dazu bemerkt werden, dass wenn zum Beispiel diese Methoden zum Information Retrieval in einem speziellen Gebiet benutzt werden sollten, dann sollte auch darauf geachtet werden, dass der Korpus sich aus einer ausreichend großen Anzahl für dieses Gebiet relevanter Dokumente aufbaut. Der Korpus in der jetzigen Form könnte z. Bsp. problemlos für journalistisches, politisches oder sportthematisches Information Retrieval genutzt werden, doch der Nutzen für ein spezielles biochemisches Dokumentensystem wäre wohl sehr fraglich.

4. Die small-world der Satzkollokationen

Wie bereits früher erwähnt, können die Satzkollokationen als ein ungerichteter Graph mit bewerteten Kanten gesehen werden. Die folgende Notation wurde frei übernommen aus Cancho & Solé [FCanSolé01].

Def.: Es sei Ω_L ein Graph der natürlichen Sprache L definiert als ein Paar $\Omega_L = (W_L, E_L)$, wobei $W_L = \{w_i\}, (i = 1,..., N_L)$ die Menge von N_L Wörtern der Sprache L ist und $E_L = \{\{w_i, w_j\}\}$ die Menge der Verbindungen zwischen Wörtern ist. Dabei bezeichnet $\xi_{ij} = \{0,1\}$ eine existierende Verbindung zwischen den Wörtern w_i und w_j und diese Wörter sind damit adjazent:

$$\xi_{ij} = \begin{cases} 1: & w_i, w_j \in W_L \quad direkt \quad verbunden \\ 0: & sonst \end{cases}$$

Bemerkung: Mit $\xi_{ij} = \{0,1\}$, einer Einschränkung auf Null oder Eins, werden die Bewertungen der Kanten, die Kollokationssignifikanzwerte, zur Vereinfachung auf ein „vorhanden“ – oder „nicht vorhanden“ reduziert, werden aber später wieder berücksichtigt.

Def.: Es sei weiterhin $\Gamma_i = \{w_j \mid \xi_{ij} = 1\}$ die Menge der mit dem Wort w_i direkt verbundenen Wörter und es gelte stets $\xi_{ii} = 0$ [i].

Bemerkungen:

- Die Menge Γ_i wird im weiteren Verlauf auch als die Nachbarn des Wortes w_i bezeichnet.
- Ohne Beschränkung der Allgemeinheit sei der Graph als zusammenhängend angenommen – bzw. es wird die größte Menge von zusammenhängenden Wörtern betrachtet und alle

[i] In dem Satzkollokationsgraphen kommen zwar „Selbstverbindungen“ vor, diese können aber bis auf weiteres ignoriert werden.

die Wörter werden ignoriert, die einzeln oder in einzelnen kleineren zusammenhängenden Gruppen stehen, damit gilt: $|E_L| \geq |W_L| - 1$

Def.: Die Dichte ϑ eines Graphen Ω_L ist eine reelle Zahl $0 \leq \vartheta \leq 1$ und bezeichnet das Verhältnis von der Anzahl der vorhandenen Verbindungen $|E_L|$ zur maximal Möglichen:

$$\vartheta = \frac{|E_L|}{\binom{|W_L|-1}{2}}$$

Der Graph der Satzkollokationen besitzt dabei folgende Eigenschaften:

- Die Anzahl der Verbindungen zwischen den Wörtern ist etwa eine Größenordnung größer, als die Anzahl der Wörter: momentan gibt es zu 1,1 Mio. Wörtern rund 16 Mio. Verbindungen[i].
- Darüber hinaus gilt auch, dass die Anzahl der benötigten Schritte, um von einem Wort ein beliebiges anderes zu erreichen sehr klein im Vergleich zur Anzahl der Knoten ist, womit es sich vermutlich bei diesem Graphen um eine „small-world“ handelt.

Sogenannte small-worlds ist eine seit kurzem, siehe Watts & Strogatz [WStrog98], untersuchte Eigenschaft von Graphen. Sie besagt unter anderem, dass small-world Graphen zwischen zufälligen und regulären Graphen liegen. Dazu werden auf einem beliebigen Graphen zwei Koeffizienten berechnet, die mittlere Weglänge d und der sogenannte clustering Koeffizient C_v. Bevor die Eigenschaft selbst definiert wird, werden an dieser Stelle zunächst reguläre und zufällige Graphen definiert.

[i] Also von 32 Mio. Kollokationseinträgen in der Datenbank die Hälfte, weil eine Verbindung durch zwei Einträge für beide Richtungen gegeben ist.

Def.: Ein regulärer Graph ist ein Tripel $\Omega_R = (P_R, V_R, <_P)$, wobei $P_R = \{k_i\}, (i = 1,..., N_R)$ die Menge von N_R Knoten ist, $V_R = \{\{v_i, v_j\}\}$ die Menge der Verbindungen zwischen Knoten und $<_P$ eine lineare Ordnungsrelation über die Elemente aus P_R. Für $i <_P j$ gilt dann für diesen Graphen, dass eine Verbindung ξ_{ij} nur dann existieren kann, also $\xi_{ij} = 1$ gilt, wenn auch $\xi_{i,(j-1)},...,\xi_{i,(i+1)}$ und $\xi_{i-1,(j-1)},...,\xi_{i-1,(i+1)}$ existieren.

Def.: Ein zufälliger Graph ist ein Paar $\Omega_Z = (P_Z, V_Z)$, wobei $P_Z = \{k_i\}, (i = 1,..., N_Z)$ die Menge von N_Z Knoten ist und $V_Z = \{\{v_i, v_j\}\}$ die Menge der Verbindungen zwischen Knoten aus P_Z bezeichnet. Für diesen Graphen gilt, dass bei einer bestimmten Dichte ϑ es gerade so viele Verbindungen zwischen zufällig ausgewählten Knoten gibt, wie benötigt werden, um die Dichte ϑ für diesen Graphen zu erreichen.

Def.: Der clustering Koeffizient C_v ist eine reelle Zahl $0 \le C_v \le 1$ und gibt über $\forall w_i \in W_L$ gemittelt an, wie oft zwei unmittelbare Nachbarn eines Wortes auch untereinander wieder verbunden sind. Dazu wird

$$L_i = \sum_{j \in \Gamma_i} \sum_{k \in \Gamma_i, j<k} \xi_{jk}$$

definiert, wobei L_i die Anzahl der Verbindungen zwischen den Nachbarn eines Wortes w_i bezeichnet. Damit lässt sich dann

$$c_v(i) = \frac{L_i}{\binom{|\Gamma_i|}{2}}$$

berechnen, welches angibt, wie viele Verbindungen zwischen den Nachbarn von w_i gegenüber der gesamt möglichen Anzahl existieren. Damit lässt sich nun der Mittelwert über alle Wörter W_L berechnen:

$$C_v = \frac{1}{N_L} \sum_{i=1}^{N_L} c_v(i)$$

Aus Dichte $\vartheta = 1$ folgt auch clustering Koeffizient $C_v = 1$, weil für jedes Wort die höchst mögliche Anzahl seiner Verbindungen existiert, womit der maximale clustering Koeffizient erreicht ist.

Def.: Die mittlere Weglänge d gibt an, wie viele Verbindungen mindestens zwischen einem beliebigen Wort w_i und einem anderen Wort w_j liegen, gemittelt über alle i und j. Dazu wird von jedem w_i ausgehend die Mindestweglänge $d_{\min}(i, j)$ zu jedem anderen Wort w_j berechnet:

$$d_v(i) = \frac{1}{N_L} \sum_{j=1}^{N_L} d_{\min}(i, j)$$

$d_v(i)$ bezeichnet damit die durchschnittliche Weglänge von einem Wort i aus, woraus der Durchschnitt d über alle Wörter aus W_L folgt:

$$d = \frac{1}{N_L} \sum_{i=1}^{N_L} d_v(i)$$

Def.: Ein Graph mit der small-world Eigenschaft ist ein Paar $\Omega_S = (P_S, V_S)$, wobei $P_S = \{k_i\}, (i = 1, ..., N_S)$ die Menge von N_S Knoten ist und $V_S = \{\{v_i, v_j\}\}$ die Menge der Verbindungen zwischen Knoten aus P_S bezeichnet. Für diesen Graphen gilt, dass für eine gewisse feste Dichte ϑ, also mit $\vartheta_S = \vartheta_Z = \vartheta_R$ bei diesem Graphen im Vergleich zu einem zufälligen Graphen der clustering Koeffizient größer ist, aber im Vergleich zu einem regulären Graphen annähernd gleich ist: $C_{v,Z} < C_{v,S} \approx C_{v,R}$. Weiterhin gilt noch, dass die mittlere Weglänge dieses Graphen im Gegenteil eher annähernd gleich der des zufälligen Graphen ist und deutlich kleiner, als die des regulären Graphen: $d_Z \approx d_S < d_R$.

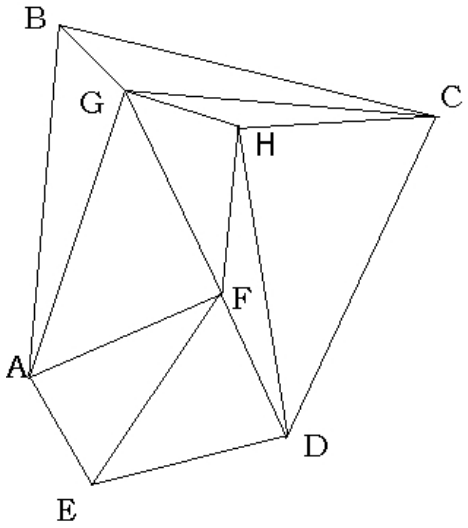
Im Zusammenhang mit dem clustering Koeffizienten ist es sinnvoll, Cluster und Cliquen zu definieren, welche lokale Maxima des clustering Koeffizienten auf dem Graphen darstellen.

Def.: Eine Clique ist ein Paar $\Omega_C = (M_C, E_C)$ und mit $\mathrm{M}_C \subseteq W_L$ ein Teilgraph von Ω_L, welcher eine Dichte von $\vartheta = 1$ hat und für den es kein $w_i \in W_L$ gibt, welches noch zusammen mit den entsprechenden Verbindungen aus E_L zu M_C hinzugenommen werden könnte und die Dichte der Clique Ω_C bei $\vartheta = 1$ gleich bleiben würde.

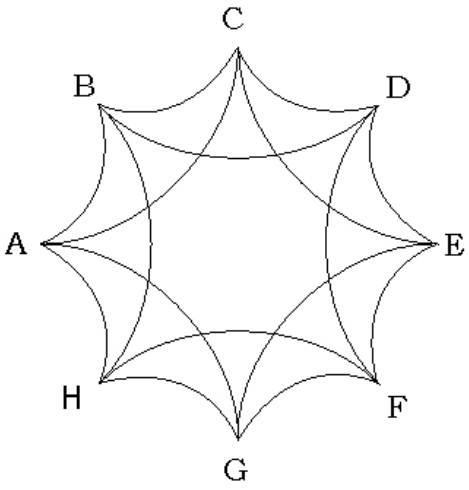
Def.: Ein Cluster ist ein Paar $\Omega_x = (M_x, E_x)$ und wie eine Clique ein Teilgraph von Ω_L, allerdings ist eine Dichte von $\vartheta \geq x$ verlangt, wobei x ein beliebiger Schwelwert $0 \leq x \leq 1$ ist und es gilt, dass es kein $w_i \in W_L$ gibt, welches noch zusammen mit den entsprechenden Verbindungen zu M_x hinzugenommen werden könnte und die Dichte der Clique M_x bei $\vartheta \geq x$ bleiben würde.

Bemerkung: Eine Clique stellt damit den Spezialfall des sogenannten maximalen Clusters dar. Außerdem bedeutet bei Graphen mit sehr kleiner Kantenanzahl ein C_v nahe 1, dass sich die vorhandenen Wörter in Clustern, bzw. Cliquen befinden müssen.

Zufälliger Graph



Regulärer Graph



Anzahl der Knoten 8
Anzahl der Kanten 16

Dichte d = 16/28 = 4/7

Abbildung 2

In der Abbildung 2 werden beispielhaft zwei Graphen der jeweiligen Art gegenübergestellt. Bei dieser Graphengröße ist weder in der Weglänge, noch im clustering Koeffizienten ein Unterschied zu bemerken, die beiden Werte entwickeln sich aber bei wachsender Graphengröße deutlich unterschiedlich, wie in Abbildung 3 sichtbar wird.

Verhalten von Weglänge und clustering Koeffizient bei gleicher Dichte und steigender Grösse

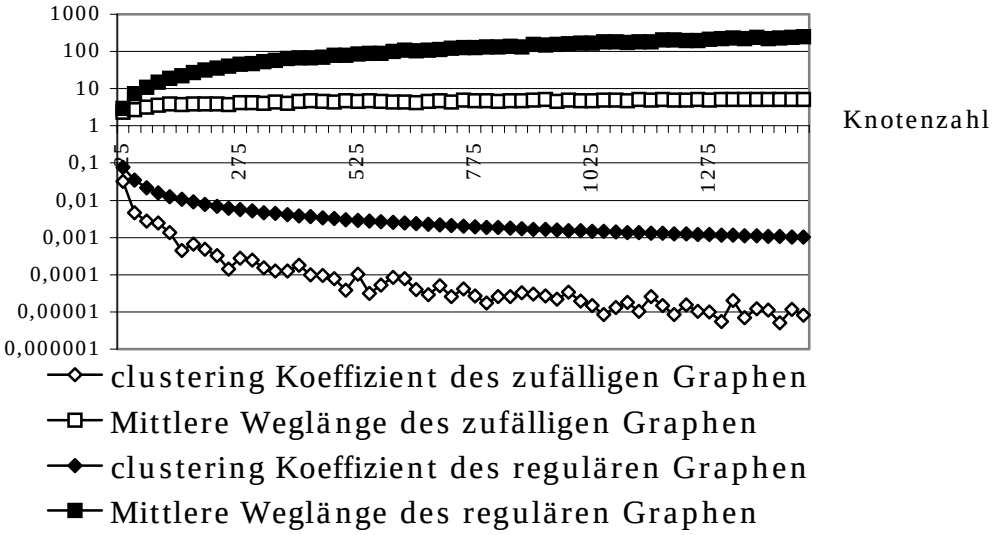


Abbildung 3

Im Vergleich zu den in Abbildung 3 gezeigten Werten stehen die stichprobenhaft gemessenen Werte des Satzkollokationsgraphen mit $C_v = 0{,}05$ (siehe Abbildung 4) und $d \approx 5$, womit deutlich wird, dass die Bedingungen $C_{v,Z} < C_{v,S} \approx C_{v,R}$ und $d_Z \approx d_S < d_R$ zumindest stichprobenhaft belegt werden können und Ω_L demnach wahrscheinlich die small-world Eigenschaft besitzt.

Die Annahme über die Existenz dieser Eigenschaften, Grundlage für die hier später vorgeschlagenen Algorithmen, lässt sich nicht ohne weiteres beweisen, sondern nur mit Beobachtungsmaterial belegen. Zum Beweisen müsste entweder per Hand der gesamte Graph durchsucht werden, oder ein NP-harter Cliquenfindalgorithmus darauf gestartet werden, der mit derzeitigen Mitteln der Rechentechnik auf dem betrachteten Graphen nicht in absehbarer Zeit terminieren würde.

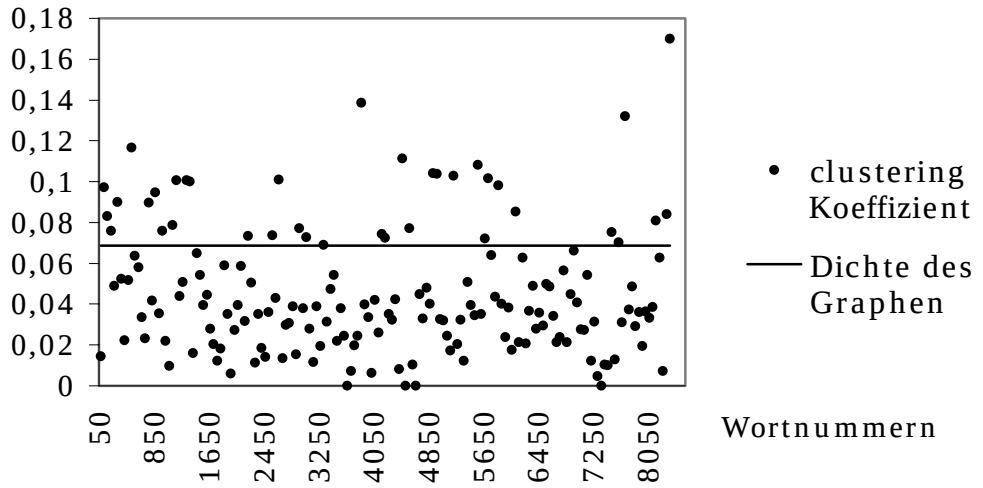


Abbildung 4

In Abbildung 4 wird die stichprobenhafte Berechnung des clustering Koeffizienten C_v für den Satzkollokationsgraphen gezeigt. Es wurden zu diesem Zweck stets 50 beliebige direkte Nachbarn eines Wortes genommen und gemessen, wie oft diese untereinander verbunden waren. Es lässt sich an dieser Stelle die Vermutung formulieren, dass der Graph folgendermaßen strukturiert ist:

- Es gibt lokale Cluster oder gar Cliquen, die über vereinzelte Verbindungen miteinander verbunden sind.
- Dabei ist jedes Wort in einem oder mehreren solcher Cluster und wenn es sich in mehreren Clustern gleichzeitig befindet, dann wird es wegen der Vermutung, dass die Cluster aufgrund der Bedeutungen der Wörter entstehen, ambiges Wort genannt. Wenn dann also 50 zufällige Nachbarn eines Wortes ausgewählt werden, kann es passieren, dass alle aus einem Cluster stammen und dann ist C_v größer, als wenn sie aus drei verschiedenen Clustern stammen, ein Hinweis darauf ist die große Streuung von C_v in dem Diagramm.

- Es gibt auch eine besondere Menge von Wörtern, die deutlich größere Nachbarschaftsmengen besitzen, als das durchschnittliche Wort und damit das verbindende Glied für weit von einander entfernte Cliquen bilden. Es handelt sich hierbei um die bereits in Kapitel 3 erwähnten Funktionswörter, also hochfrequente aber wenig inhaltstragende Wörter, die in jedem beliebigen Satz unabhängig des Inhalts auftreten können, weil sie nur funktionale Aufgaben besitzen. Diese werden von hier an long-range Wörter genannt.

In der Abbildung 5 ist diese Struktur schematisch dargestellt.

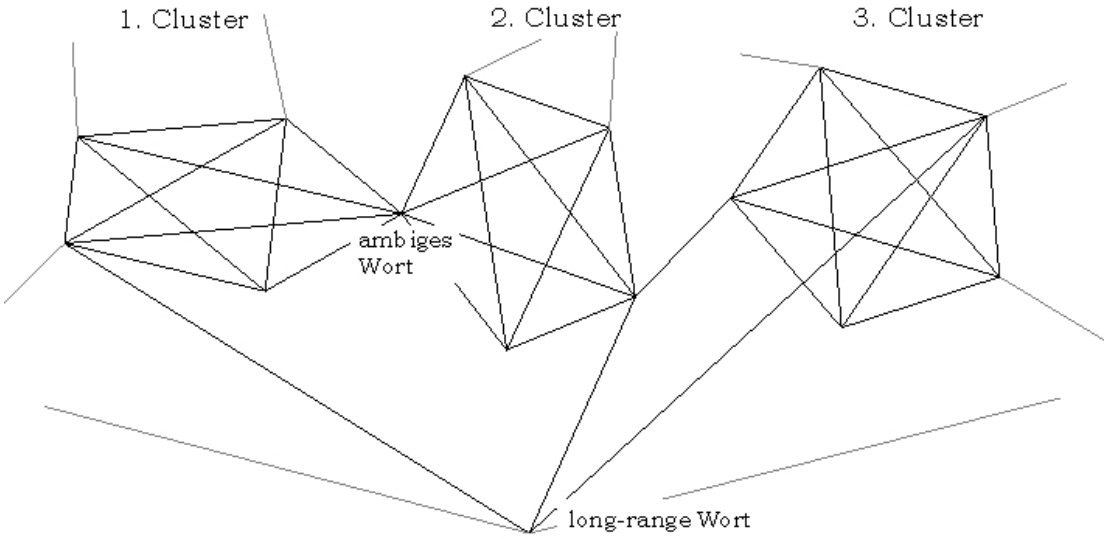


Abbildung 5

Zu den long-range Wörtern ist noch zu erwähnen, dass die small-world Eigenschaft eines Graphen ursprünglich beispielhaft an Daten über amerikanische Filmschauspieler gezeigt wurde. Es ging darum, dass ein Schauspieler direkt mit einem anderen assoziiert war, wenn er zusammen mit diesem anderen in einem Film gespielt hatte. Dieser Graph hatte gerade die Eigenschaften, wie sie weiter oben beschrieben wurden, bis auf das Analogon zu den long-range Wörtern – es gab also keine Schauspieler, die mit sehr vielen anderen Schauspielern in sehr vielen Filmen gespielt hätten. Später wurde die small-world

Eigenschaft allerdings an ganz anderen Stellen gezeigt, wie etwa dem Stromnetz eines Landes, welches wiederum auf jeden Fall derartige long-range Knoten hatte – die Hochspannungsleitungen zwischen den Städten.

Es gibt also möglicherweise zwei verschiedene Arten der small-world Eigenschaft, wobei beide allerdings die für diese Arbeit wichtige Gemeinsamkeit besitzen, dass es lokale Cluster gibt. Die long-range Wörter werden durch Frequenzfilter weggefiltert.

4.1. Bedeutung der Cluster

Die Cluster eines Kollokationsgraphen selbst wiederum sind in Abhängigkeit von der zugrundeliegenden Kollokationsart unterschiedlichen Inhalts. Bei Nachbarschaftskollokationen lassen sich grammatisch passende und gleichzeitig semantisch verwandte Wörter für Substantive im gleichen Cluster beobachten (als Beispiel hier das Wort Führung):

> politische (714), unternehmerische (500), politischen (406), chinesische (381), irakische (311), russische (287), iranische (286), serbische (274), Pekinger (269), chinesischen (202), militärische (175), Belgrader (174), kommunistische (174), guter (166), iranischen (156), Chinas (148), industrielle (145), Unter (144), russischen (134), irakischen (125), serbischen (123), Moskauer (111), sowjetische (103), Teheraner (100), tschetschenische (90), militärischen (88)

Bei Satzkollokationen hingegen finden sich die Cliquen offenbar auf nahezu rein semantischer Ebene:

> Ausstellung (782), Minute (772), übernahm (613), politische (473), Volkshochschule (401), Sonderausstellung (386), Unter (381), unternehmerische (362), Minuten (325), Prähistorische (303), Konsortium (297), Museum (289), übernehmen (266), Staatssammlung (264), politischen (260), brachte (255), Bankenkonsortium (251), Zuschauern (245), Neue Pinakothek (242), chinesische (241), bosnischen (240), Partei (234), serbische (230), Kunst (224), Gesamtwertung (223), baute (211), iranische (211)

Es ist erkennbar, dass hier einige Bedeutungen gemischt auftreten, besser lässt sich das bei dem visualisierten Teilgraphen erkennen:

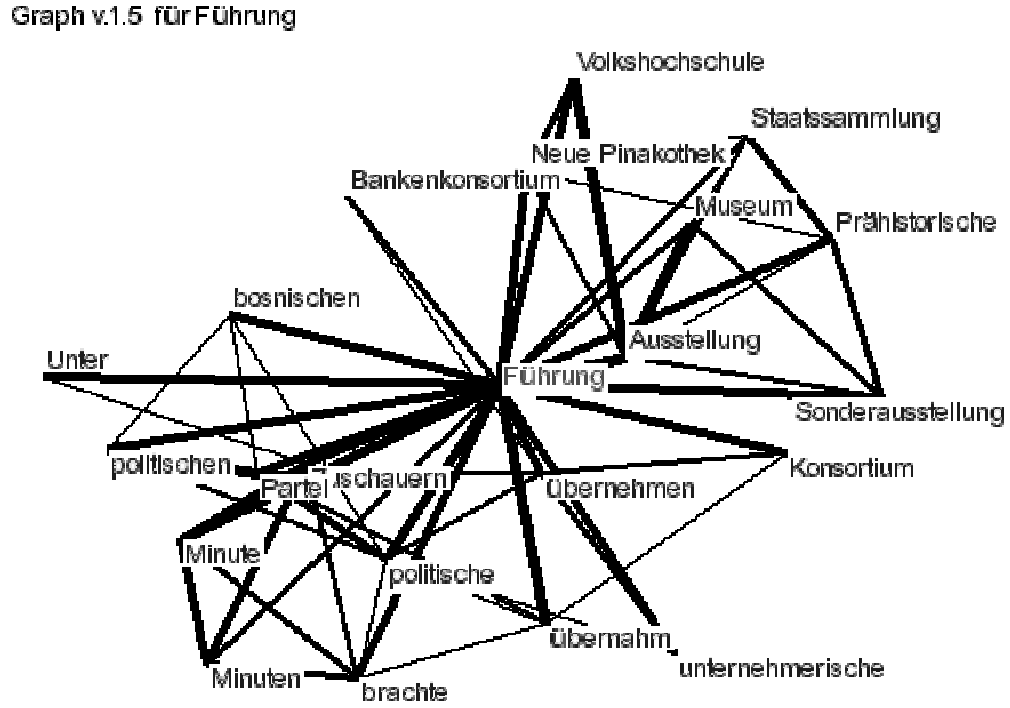


Abbildung 6

Bei diesen Visualisierungen von Kollokationsbeziehungen wurden zu einem Eingabewort diejenigen Kollokate eingezeichnet, die auch untereinander signifikante Beziehungen besitzen, siehe auch F. Schmidts Diplomarbeit [Schmidt99]. Die optimale Variante wäre die, einen Algorithmus zu haben, der alle Cluster findet, in denen sich das Wort befindet. Ein solcher vollständiger Algorithmus könnte wie bereits erwähnt, nicht mit vertretbaren Aufwand angewendet werden und daher ist es wichtig, eine günstige Approximation für selbigen zu finden. Diese Approximation sollte die beteiligten Cluster sauber genug trennen können (Precision) und auch genügend entsprechende Knoten liefern (Recall). Dabei sollte das Problem beachtet werden, dass einige Cluster, in denen sich ein Wort befindet, wesentlich mehr Wörter beinhalten, als andere.

Eine derartige Approximation würde eine solche Dichotomie liefern, wie sie früher verlangt wurde, um die gemischte Kollokationsmenge sinnvoll einzuteilen, und wird im folgenden Kapitel beschrieben.

5. Clusterapproximation durch Tripel

Wie im vorigen Kapitel bereits ausführlich beschrieben, ist das Kernelement des Vorgehens in dieser Arbeit der Fakt, dass in dem Kollokationsgraphen Assoziationen, bzw. Verbindungen zwischen einzelnen Wörtern gegeben sind. Ein wichtiges Problem mit diesen Verbindungen ist es, dass unspezifiziert ist, welche Relation sie zwischen den beiden verbundenen Wörtern angeben. Weiterhin gibt es die Sachgebietsdatenbank, in der Wörter Sachgebieten zugeordnet sind.

Semantische Eigenschaft eines Knotens ist eine beliebige Zuweisung von externem Inhalt zu einem Knoten. Ein Beispiel einer derartigen Zuweisung ist die Zuweisung eines Wortes zu einem Sachgebiet.

Assoziationen zwischen Wörtern eines Graphen sind Beziehungen unspezifizierter Art zwischen den Wörtern, wie zum Beispiel die Beziehung der Kollokation zwischen zwei Wörtern bzw. der Verbindung zweier Wörter in dem Graphen.

Damit soll auch unter anderem deutlich werden, dass die hier vorgeschlagenen Vorgehensweisen nicht auf die konkrete Datensituation beschränkt sein müssen, in der sie entwickelt wurden. Der Terminus „Wort“ wird hier synonym zu „Knoten“ benutzt. Eigenschaftslose Wörter sind die Wörter, die keine Eigenschaft besitzen, bzw. bei denen die Eigenschaften unbekannt sind oder denen noch keine Eigenschaften zugewiesen wurden.

Ziel ist also unter den gegebenen Umständen, allen Wörtern, die keine Eigenschaften besitzen, bzw. die sie wohl besitzen, aber bei denen sie unbekannt sind, anhand der Assoziationen zu anderen Wörtern, bei den die Eigenschaften bekannt sind, ebenfalls Eigenschaften zuzuweisen. Dabei sollte diese Zuweisung möglichst dem nahe kommen, wie auch ein Mensch sie zugewiesen hätte. Einstellbar sollte auch sein, ob ein Knoten nur einmal eine Eigenschaft haben kann

oder auch mehrere, um damit die verschiedenen Module wie in Kapitel 2 erwähnt, anbieten zu können.

5.1. Direkte iterative Vererbung

Vererbung von semantischen Eigenschaften ist eine Operation, bei der einem Wort eine semantische Eigenschaft aufgrund seiner Verbindungen zu anderen Wörtern zugewiesen wird.

Die naive Herangehensweise ist die, dass ein eigenschaftsloses Wort, das über mindestens eine direkte Verbindung zu einem anderen Wort mit einer Eigenschaft verfügt, diese Eigenschaft erbt. Im Zweifelsfalle, also wenn sich mehrere solche Wörter in der direkten Nachbarschaft befinden, kann die am meisten vorkommende Eigenschaft ausgezählt werden. Dies könnte direkte Nachbarschaftsvererbung der Eigenschaft genannt werden, und dieses Verfahren lässt sich für den speziellen Fall der Wortschatz-Datenbank sogar noch um einiges verfeinern, je mehr Faktoren bei der Berechnung mit eingeschlossen werden. So kann und sollte beachtet werden:

- Ob die Verbindungen gewichtet sind (Kollokationssignifikanz wäre eine solche Gewichtung).
- Ob und in welcher Richtung sich die Frequenzen der beteiligten Wörter unterscheiden.
- Wie viele der Verbindungen dieses Knotens insgesamt überhaupt zu Wörtern mit Eigenschaften führen (ob das Wort sich in einem sehr unbekannten Gebiet befindet oder ringsherum bereits alles bekannt ist).
- Wie stark die bei der Auszählung beteiligten Eigenschaften selbst in der gesamten Datenmenge vertreten sind. (Wenn es eine Eigenschaft gibt, die nur einem einzigen Knoten zugewiesen wurde, so wird sie mit großer Gefahr jede Auszählung verlieren).
- Wie groß die Anzahl der überhaupt vorhandenen Eigenschaften ist.

Abhängig von der Art der dem Problem zugrundeliegenden Daten können noch eine Vielzahl anderer Faktoren hinzukommen.

Weiterhin kann hier zwischen bottom-up oder top-down Verfahren unterschieden werden. Bottom-up kann man unformal so beschreiben, dass von eigenschaftslosen Knoten ausgegangen und wird ein Verfahren angewendet wird, um diesem Knoten eine Eigenschaft zuzuweisen. Top-down wäre hingegen, dass von einem Knoten mit bekannter Eigenschaft ausgegangen wird und versucht wird, allen Nachbarn diese Eigenschaft zu vererben.

Das gesamte Verfahren könnte schließlich so lange iteriert werden, bis alle Knoten eine Eigenschaft besitzen würden, doch würden sich bereits nach einige Durchläufen Fehlzuweisungen fatal auswirken.

Der Unterschied zwischen diesen Verfahren äußert sich vor allem darin, wofür sie eingesetzt werden sollen. Will der Benutzer des Systems lediglich für einen Knoten die passendste Eigenschaft wissen, sollte bottom-up verwendet werden. Sollte es aber der Fall sein, dass die Auswirkung einer Art von Eigenschaften in dem Graphen untersucht werden soll, dann ist die top-down Methode die günstigere, da sonst jeweils der gesamte Graph durchgerechnet werden müsste, bevor das gewünschte Ergebnis erreicht wird. Dies ist also die Stelle, an der sich dieses Verfahren den Wünschen des Benutzer leicht lässt.

5.2. Ausnutzung zugrundeliegender Strukturen

Der Nachteil des beschriebenen Verfahrens ist der, dass es die dem Graphen immanente Struktur nur sehr begrenzt ausnutzt. Wie im vorigen Kapitel beschrieben, lassen sich in dem Graphen Cluster beobachten, die semantisch bedingt zu sein scheinen. Eine direkte Vererbung würde allerdings den Einfluss dieser Cluster ignorieren und damit eventuell deutlich schlechtere Ergebnisse erbringen, als möglich wäre. Angenommen, diese Cluster bedeuteten wirklich, dass alle an einem Cluster beteiligten Knoten auch semantisch zusammengehörten – Evidenz für diese Annahme, siehe Fabians Diplomarbeit

[Schmidt99] (Kap. 3.3 S. 41). Dann sollte also die in dieser Clique am häufigsten vorkommende Eigenschaft auf die eigenschaftslosen Mitglieder mit einem Schritt vererbt, ein besseres Ergebnis bringen, als wenn iterativ direkt vererbt wird, bis jeder Knoten eine Eigenschaft hat – dabei würden zwar auch, bedingt durch die Struktur, Mitglieder von Clustern gegenüber nicht-Mitgliedern von Clustern bevorteilt werden, allerdings wäre das lediglich ein Nebeneffekt und könnte, weil nicht voll ausgenutzt, dazu führen, dass andere Cluster ohne weiteres mitvererbt werden.

In dem Fall also, dass diese Annahme stimmt, dann ist die Vermutung, dass wenn ein Algorithmus lediglich alle möglichen Cluster findet, und für jeden bestimmt wird, welche Eigenschaft darin jeweils am stärksten vertreten ist und diese dann auf den Rest des jeweiligen Clusters vererbt, dass das Ergebnis ein besseres ist, als mit direkter iterativer Vererbung. Ob diese Vermutung stimmt, hängt natürlich nicht nur von der Annahme über die Semantizität der Cluster ab, sondern von der tatsächlichen Existenz der im vorigen Kapitel angenommen Cluster und dass diese Cluster sich in der Tat durch die Bedeutung der Wörter zusammenfinden. Wenn konkret bei der Wortschatz-Datenbank die der Sachgebietszuweisung zugrunde liegende natürliche Einteilung zu sehr von dem abweicht, was die Cluster hier zusammenfindet, dann können auch keine guten Ergebnisse erwartet werden.

Dennoch wären die Ergebnisse einer solchen Clustersuche selbst nicht umsonst, denn sie würden lediglich eine andere Einteilung der Daten bedeuten, als die gewollte und könnten für unvorhergesehene andere Zwecke genutzt werden – sie würden eine natürliche Klassifizierung der Daten darstellen, wie später zu sehen sein wird.

Maximale Cluster in einem Graphen zu finden ist allerdings ein NP-vollständiges Problem und kann nicht einmal auf einem Bruchteil der Daten in vernünftiger Zeit durchgeführt werden und daher wird eine Approximation benötigt, die Cluster möglichst gut findet, aber nicht

so berechnungskomplex ist. Ausserdem ist es für die gegebene Aufgabe nicht wichtig, alle Cluster mit einem Mal zu finden - es reicht, wenn von einem Knoten ausgehend alle angrenzenden Cluster gefunden werden. Zusätzlich müssen hier wahrscheinlich nicht alle möglichen komplexen Arten von Clustern gefunden und getrennt werden – es reicht eventuell bereits eine gröbere Einteilung.

5.3. Approximation durch Tripel

Dieser Lösungsansatz untersucht Tripel von Wörtern in ihrem möglichen Zusammenhang mit ihrer unmittelbaren Umgebung. Es wird zu einem Tripel von Wörtern die Menge von Wörtern gesucht, die Verbindungen zu allen drei Wörtern dieses Tripels besitzen.

Def.: Die Berechnung eines charakteristischen Vektors ist eine Abbildung $X:\{w_i, w_j, w_k\} \rightarrow \upsilon_{ijk}$, wobei $\upsilon_{ijk} = \{w_1, w_2, ..., w_{N_L}\}$ mit $w_i = \{0 \mid 1\}$ und es gilt:

$$\upsilon_{ijk}\{l\} = \begin{cases} 1: & w_l \in \Gamma_i \wedge w_l \in \Gamma_j \wedge w_l \in \Gamma_k \\ 0: & sonst \end{cases}$$

Der charakteristische Vektor υ_{ijk} eines Tripels von Wörtern $w_i, w_j, w_k \in W_L$ ist also eine Menge von Wörtern, welche aufgrund ihrer Position im Graphen Ω_L relativ zu den Wörtern des Tripels als charakteristisch für dieses Tripel betrachtet werden. Dabei wird ein Wort w_l in diesem Fall genau dann eine 1 an seiner Stelle l in υ_{ijk} haben, wenn es in allen drei Mengen Γ_i, Γ_j und Γ_k gleichzeitig ist, bzw. die Schnittmenge der drei Mengen enthält alle Wörter, die in υ_{ijk} eine 1 haben.

Wenn w_i, w_j und w_k Elemente einer Clique M_L sind, so wird der charakteristische Vektor mindestens sämtliche Knoten der Clique beinhalten, da gilt: $\forall w_l \in M_L [w_l \in \Gamma_i \wedge w_l \in \Gamma_j \wedge w_l \in \Gamma_k]$

Die Definition eines charakteristischen Vektors ist problemlos auf beliebig große Tupel erweiterbar:

$$\upsilon_{ijk}\{l\} = \begin{cases} 1: & w_l \in \Gamma_i \wedge w_l \in \Gamma_{i+1} \wedge \ldots \wedge w_l \in \Gamma_{i+n} \\ 0: & \textit{sonst} \end{cases}$$

Je größer allerdings das Tupel, umso größer auch die Mindestgröße der vermuteten Clique. Hier werden bevorzugt Tripel verwendet, weil die Wahl der Größe der Tupel von der Gesamtdichte des Graphen abhängt und Tripel für den Satzkollokationsgraphen, wie zu sehen sein wird, am günstigsten waren. Für wesentlich dichtere Graphen oder einige besonders dichte Stellen in dem betrachteten Graphen könnten Quadrupel besser sein, weil bei steigender Dichte die entsprechenden Cluster größer werden und die Zahl der Wörter wächst, die in der Schnittmenge auftreten, also die Bedingung erfüllen, zu allen Eingabewörtern assoziiert zu sein, obwohl sie nicht zum Cluster gehören.

Es könnte aber auch über die Clique hinaus Wörter geben, die mit allen Elementen aus dem Tripel $w_i, w_j, w_k \in W_L$ verbunden sind, obwohl nicht mit allen Elementen aus M_L. Um zu prüfen, ob das nächste betrachtete Wort ein Wort der Clique ist, müssten alle seine möglichen Verbindungen zu bereits akzeptierten Knoten der Clique geprüft werden. Das ist der Punkt, an dem die Approximation aussagt, dass es in diesem Graphen wenig wahrscheinlich ist, dass Knoten diese Eigenschaften erfüllen, ohne auch zur Clique bzw. zum Cluster zu gehören und es daher nicht prüft.

Es handelt sich allerdings lediglich in seltenen Fällen um echte Cliquen als Teilgraphen, es sind vielmehr Cluster mit einer Dichte von ca. $\vartheta = 0{,}5$.

Demnach wird der Schnitt der direkten Nachbarn eines Tripels in solchen Fällen lediglich eine Teilmenge eines zugrundeliegenden Clusters beinhalten. Je geringer die Dichte des Clusters, umso größer wird die Wahrscheinlichkeit, dass die Schnittmenge leer bleiben wird. Das

ist allerdings auch gleichzeitig dasjenige Kriterium, welches Cluster von nicht-Clustern unterscheidet: Liefert der Schnitt die leere Menge, so wird angenommen, dass die drei Eingabeknoten nicht aus einem gemeinsamen Cluster stammen.

Da der betrachtete Satzkollokationsgraph aber kein „normaler“ clusternder Graph ist, sondern einer mit der small-world Eigenschaft, gibt es einige Wörter – die long-range Wörter, in deren direkter Nachbarschaft sich große Teile des Gesamtgraphen befinden. Es gibt demnach eine hinreichend große Wahrscheinlichkeit, dass der Schnitt der Assoziationsmengen dreier beliebiger Knoten ein oder mehr solcher long-range Wörter beinhaltet, was zu der evtl. falschen Annahme verleiten könnte, sie stammten aus dem gleichen Cluster. Dieses Problem lässt sich auf zweierlei Weisen umgehen:

1. Die long-range Wörter werden als solche identifiziert und ignoriert (wie es in dieser Arbeit auch getan wird: Eine einfache Approximation war es, alle Wörter mit einer Frequenz größer als ein bestimmter Wert zu ignorieren, womit damit effektiv etwa die häufigsten 1000 Wörter im Wortschatz-Lexikon betroffen waren)
2. Es werden zusätzlich Zwischenbeziehungen zwischen den Eingabeknoten, sowie den gefundenen Knoten überprüft. Es wird zum Beispiel die Dichte des gefundenen als Cluster vermuteten Teilgraphen berechnet und sollte diese kleiner als ein Schwellwert sein, müssten die Ergebnisse abgelehnt werden.

Allgemein könnte es aber in der Zukunft von Interesse sein, gerade die long-range Wörter selbst zu untersuchen, mit welchen Clustern sie bevorzugt verbunden sind, mit welchen eher nicht usw., an dieser Stelle konnte kein direkter Nutzen aus ihnen gezogen werden.

5.4. Erweiterte charakteristische Vektoren

An besonders dünnen Stellen in einem Graphen[i], wird die bisher diskutierte Definition eines charakteristischen Vektors oft die leere Menge liefern, obwohl Cluster zu erkennen sind. Allerdings handelt es sich dann eher um Cluster, deren Dichte sehr gering ist, aber immer noch signifikant größer als wenn die umliegenden Knoten mit hinzugerechnet werden würden. An dieser Stelle ist es sinnvoll, eine erweiterte Berechnung des charakteristischen Vektors zu benutzen, die eine schwächere Bedingung stellt.

Hierzu wird zunächst eine Zwischenmenge von Tripeln generiert, die erweiterte Eingabemenge. Die folgenden Definitionen lassen sich wieder problemlos auf beliebige Tupel erweitern, wovon aber der Einfachheit halber abgesehen wurde.

Def.: Die erweiterte Eingabemenge $B = \{\tau_1, \tau_2, ..., \tau_n\}$ mit $\tau_l = \{w_m, w_n, w_o\}$ für beliebige $w_m, w_n, w_o \in W_L$ ist eine Menge von Tripeln, die nach einer Vorschrift aus dem ursprünglichen Eingabetripel $T = \{w_p, w_q, w_r\}$ aufgrund der Verbindungen der Wörter untereinander generiert wurde. Es wird ein Tripel B_x in B aufgenommen, wenn B_x folgende Bedingungen erfüllt:

$$\forall w_i w_j \in T \forall w_k w_l \in B_x [w_k \in \Gamma_i \wedge w_l \in \Gamma_i \wedge w_j \in \Gamma_i]$$

Def.: Der erweiterte charakteristische Vektor ω_{ijk} ist an dieser Stelle dann folgendermaßen definiert, wobei υ_{τ_i} der charakteristische Vektor des Tripels $\tau_i \in B$ ist:

$$\omega_{ijk} = \upsilon_{\tau_1} + \upsilon_{\tau_2} + ... + \upsilon_{\tau_n}$$

Es wird also eine Menge von Tripeln generiert, bei denen dann jeweils wieder die charakteristischen Vektor ausgerechnet werden. Die Menge der charakteristischen Vektoren werden zum Schluss zu einem

---

[i] Ein Teilgraph ist dünn, wenn seine Dichte ϑ einen kleinen Wert hat.

neuen Gesamtvektor addiert, wobei jede Zahl größer Null zur Vereinfachung auf Eins normiert wird.

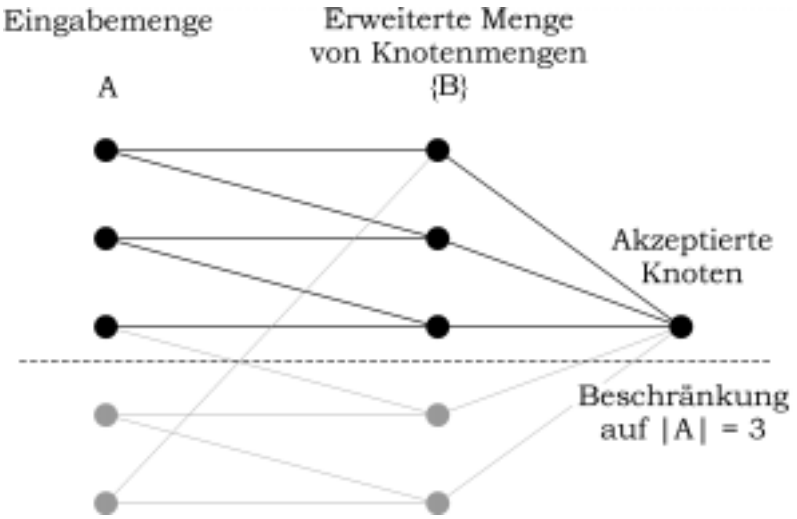


Abbildung 7

Diese Definition erfüllt nicht die Bedingungen, wie sie mit der direkten Schnittmenge gegeben war:

- Eine gelieferte Menge von Knoten, die eventuell ein Cluster darstellt, muss nicht die Eingebewörter beinhalten.
- Es können mehrere Cluster gemischt gefunden werden, obwohl die Eingabewörter aus einem einzigen Cluster stammen.
- Es können Cluster gefunden werden, obwohl die Eingabewörter nicht alle in einem Cluster lagen.
- Auf normale Tripel angewendet, also solche, bei denen ohnehin bereits ein Cluster gefunden worden war, bzw. der charakteristische Vektor nicht nur Nullen enthielt, ergibt dieser Algorithmus eine wesentlich vergrößerte Menge, eine Art „Übercluster“, in dem unter anderem auch der bereits früher Gefundene mit enthalten ist.

Bei dieser Definition werden eine Reihe Annahmen über Existenzen von Verbindungen getroffen, die auf Kosten der Laufzeit zusätzlich

überprüft werden könnten, um die Präzision der Ergebnisse zu erhöhen. Sie sind in folgender Abbildung erkennbar:

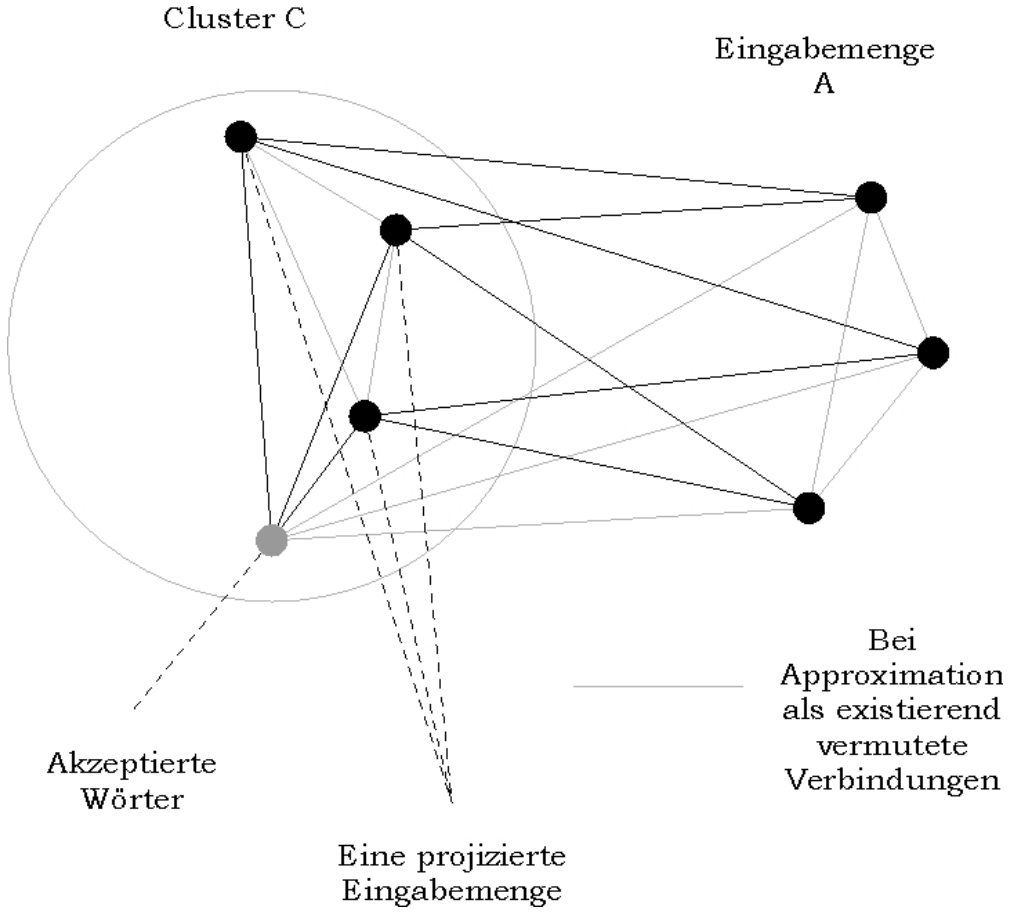


Abbildung 8

In jedem Falle (bzw. bei einer sinnvollen Definition des charakteristischen Vektors) wird ein Termvektor zurückgeliefert, der charakteristisch für dieses Tripel ist. Meistens wird dieser Termvektor keine Terme enthalten, denn die meisten Worttripel liegen zu weit voneinander entfernt. Für alle Tripel, die nahe genug beieinander liegen, wird es jeweils unterschiedliche Termvektoren geben, außer in den seltenen Fällen, dass zwei verschiedene erweiterte Eingabetripel B_x und B_y aus ein- und der gleichen echten Clique stammen und diese Clique ansonsten scharf genug vom Rest des Graphen abgegrenzt ist.

Das Resultat dieser Approximation in der betrachteten Art von Graphen (siehe voriges Kapitel) ist wie folgt (wenn die Filterung der long-range Wörter nicht durchgeführt wird):

1. die leere Menge, wenn nicht alle drei Wörter des Eingabetripels aus der gleichen Clique stammen
2. eine Menge von long-range Wörter, wenn zwar nicht wie bei 1. alle Wörter aus der gleichen Clique stammen aber zufällig über das gleiche long-range Wort verbunden sind
3. eine Menge von Wörtern aus der Clique, sowie eventuell eine Menge von long-range Wörtern, wenn alle drei Wörter des Eingabetripels aus der gleichen Clique stammen.

5.5. Ergebnisse

Im Weiteren werden einige beispielhafte Berechnungen von Tripeln nach der trivialen Definition aufgeführt. Dabei wird zunächst ein Beispiel eines echt ambigen Tripels gezeigt, bei welchem die Ergebnismenge dann einen entsprechend geringen clustering Koeffizienten c aufweist, während es mindestens eine Einteilung dieser Menge in Teilmengen gibt, deren clustering Koeffizienten jeweils deutlich höher liegen:

$C_v = 0.2094$: Gold Silber Bronze

100Meter 200Meter Abfahrt Albertville Almsick Aluminium Atlanta ausgezeichnet Barcelona Barren Dagmar Dreimal dreimal Edelmetall Edelmetallen Einzel Eisen Elfenbein EM erfolgreichste erkämpfte errang Europameisterschaften Finale Franziska Freistil fünfmal geholt gewann gewannen gewonnen GUS Göteborg Hase holte holten Kanada Keramik kg Kupfer Legierung Lillehammer m Mannschaftswettbewerb Marmor Medaille Medaillen Mehrkampf Messing Metall Nagano Nationenwertung neunmal Olympia Olympiasieger Olympischen olympischen Rettungsschwimmerabzeichen Riesenslalom Schmuck sechsmal Seoul sicherte siebenmal Spielen Staffel Stoßen Super-G Südkorea Team v.Chr. van viermal Wasserwacht Weitsprung Weltmeister Weltmeisterschaft Weltmeisterschaften Winterspielen WM Zinn Zweimal zweimal

$C_v = 0.3841$: Gold Silber Kupfer

abgebaut Aluminium Antimon Blei Bodenschätze Bodenschätzen Bronze Chrom Diamanten Edelmetalle Edelmetallen Edelsteine Eisen Eisenerz Erdöl Erze Erzen gewonnen Gußeisen Kalkstein Kilogramm Kohle Mangan Marmor Mengen Messing Metall Metalle Metallen Mineralien Molybdän Nickel Palladium Platin Quecksilber reich Salz Schwefel Tonnen Uran v.Chr. Vanadium Vorkommen Wolfram Zink Zinn

Es ist hier deutlich zu sehen, wie das Finden des Maximums des clustering Koeffizienten c auch die Qualität der gefundenen Mengen verbessert. Dazu reichte es bereits aus, eines der drei Eingabewörter durch eines der Wörter in der Ausgabemenge zu ersetzen und damit ausnutzen, dass die Schnittmengen zwischen zwei Clustern weit kleiner sein müssten als die Cluster selbst. In diesem Falle wurde Bronze durch Kupfer ersetzt. Danach wurden alle Wörter, die für dieses neue Tripel gefunden wurden entfernt und die restlichen falschen markiert:

100Meter 200Meter Abfahrt Albertville Almsick Atlanta ausgezeichnet Barcelona Barren Dagmar Dreimal dreimal Edelmetall Einzel Elfenbein EM erfolgreichste erkämpfte errang Europameisterschaften Finale Franziska Freistil fünfmal geholt gewann gewannen gewonnen GUS Göteborg Hase holte holten Kanada Keramik kg Legierung Lillehammer m Mannschaftswettbewerb Medaille Medaillen Mehrkampf Nagano Nationenwertung neunmal Olympia Olympiasieger Olympischen olympischen Rettungsschwimmerabzeichen Riesenslalom Schmuck sechsmal Seoul sicherte siebenmal Spielen Staffel Stoßen Super-G Südkorea Team v.Chr. van viermal Wasserwacht Weitsprung Weltmeister Weltmeisterschaft Weltmeisterschaften Winterspielen WM Zweimal zweimal

Mit einem Schritt wurde demnach die ursprüngliche Menge der Wörter sinnvoll in verschiedene Mengen zerlegt, die bis auf einige Ausnahmen bedeutungshomogen sind. Statt dessen hätte auch ein aufwändiger Algorithmus gestartet werden können, welcher zwar die genauen Teilmengen gefunden hätte, aber auch wesentlich länger gebraucht hätte, bei unwesentlicher Verbesserung der Precision – von 76 Wörtern sind nur noch vier falsch. Darüber hinaus ist das Ausrechnen eines neuen Tripels eigentlich keine korrekte Maximumfindung über der Wortmenge des charakteristischen Vektors von ‘Gold, Silber, Bronze’, da sie auch neue Wörter mitliefert, die von dem ursprünglichen Tripel nicht erreichbar waren. Diese Eigenschaft wiederum ist ein erwünschter Nebeneffekt ist – er erhöht den Recall, ohne an Precision zu verlieren, außer in dem unwahrscheinlichen Fall, mit dem modifizierten Tripel in eine Schnittmenge hineinzugeraten, welche noch ein drittes Cluster involviert.

Dennoch gibt es noch größere Maxima des clustering Koeffizienten, wenn wiederum andere Tripel benutzt würden, die eine noch genauere Teilmenge der Edelmetalle herausgreift:

$\underline{C_v = 0.8356}$: Blei Uran Chrom

Arsen Eisen Eisenerz Gold Kohle Kupfer Mangan Nickel Silber Tonnen Wolfram Zink Zinn

Von einem niedrigen clustering Koeffizienten lässt sich allerdings leider nicht ableiten, dass in der gefundenen Menge mehr als ein lokales Maximum existiert oder dass es sich um Wörter verschiedener Bedeutungen handelt:

$\underline{C_v = 0.0817}$: Moslems Hindus Christen

Ausschreitungen beten blutigen Buddhisten gläubige heiligen islamische Moschee religiösen Sikhs Zusammenstößen

In diesem Fall resultiert der niedrige Koeffizient eher daher, dass es in diesem Bereich des Graphen insgesamt wenig Verbindungen gibt, sich die gefundene Teilmenge aber dennoch von der Umgebung abhebt.

Weiterhin gilt, dass je mehr Knoten eines Graphen mit einer festen Anzahl von Kanten genommen werden, umso geringer die Wahrscheinlichkeit, einen bestimmten clustering Koeffizienten zu erreichen. Dennoch können die Termvektoren recht umfangreich werden und dennoch einen hohes clustering Koeffizienten zu besitzen, was ein deutliches Indiz dafür ist, dass es sich in der Tat um ein Cluster handeln muss:

$\underline{C_v = 0.4826}$: Stich Viertelfinale Achtelfinale

Agassi Andre ausgeschieden Australian Australier Becker Bernd besiegte bezwang Boris Cedric Courier David Doppel dotierten Dreekmann Duell Edberg Einzel Ferreira Finale French Gegner gesetzte gewann gewonnen Goran Graf Haas Halbfinale Hendrik Huber Ivanisevic Jewgeni Jim Kafelnikow Karbacher Korda Krajicek Leimener Mailand Marc Match McEnroe Melbourne Michael Muster Niederlage Niederländer Nummer Open Ostrau Patrick Pete Petr Pioline Prinosil Qualifikanten Richard Rosset

Rothenbaum Runde Sampras Schweden Steeb Stefan Stich Sätzen Team Tennis Tennisturnier Tennisturniers Thomas trifft Tschechen Turnier Turniers unterlag US verlor Wayne Weltrangliste Wimbledon Wimbledonsieger zog

Im Vergleich dazu eine Ergebnismenge etwa des gleichen Umfanges, bei welchem allerdings ein sehr kleiner clustering Koeffizient gemessen wurde:

$\underline{C_v = 0.0318}$: Monate Stunden Tage

Ablauf Acht anderthalb Anzahl Arbeitstage Aufenthalt Ausbruch Bereits beträgt Binnen binnen brauchte brauchten Dauer dauere dauern dauernde dauernden dauert dauerte dauerten Drei Durchschnitt durchschnittlich durchschnittliche ehe eineinhalb eingesperrt Einige elf Entführung festgehalten Festnahme folgenden freigelassen Frist Fünf gearbeitet gedauert gesperrt gewartet hielt hinweg höchstens Innerhalb Klinik Krankenhaus länger maximal Monat Neun Operation Pause saß Schnitt schweren Sechs Sieben spätestens starb Start Tat unterwegs Urlaub verbracht verbrachte verbringen verbringt vergangen vergehen vergingen Verhaftung verkürzen verkürzt verlängern verlängert Verlängerung Vier Vierzehn vierzehn vorher warten Wartezeit Wenige Wohnung währenden währte Zehn Zeitraum zubringen zugebracht Zweieinhalb zweieinhalb zweimal Zwölf

Die drei Eingabewörter hier haben eine hohe Frequenz und tragen wenig spezifischen Inhalt. Wie man sieht, können diese Wörter allesamt in beliebigen Kontexten verwendet werden, sie ähneln damit eher Funktionswörtern.

Es werden im folgenden einige weitere beispielhafte Berechnungen von Tripeln angeführt, es wird unter anderem demonstriert, wie sowohl ein weiterer Kontext für ‘Gold’ gefunden werden kann, als auch Redewendungen. Weiterhin wird mit drei vollkommen verschiedenen Tripeln ein Cluster von verschiedenen Stellen beleuchtet. Die charakteristischen Vektoren dieser Tripel enthalten dann einige gleiche Wörter:

w_1, w_2, w_3	$v_{1,2,3}$	C_v
Gold wurde Preis	ausgezeichnet Auszeichnung Award erhielt Feinunze Fixing gewann Verdienste verkauft verliehen	0.3731
Schiff verlassen sinkende	Ratten	0.0
Courier Runde Steeb	Achtelfinale Becker Einzel gewann GRAND-PRIX-TURNIER Match Michael Stich Tennis Turnier Viertelfinale	0.8310

Chang Court Doppel	Becker Boris Becker French Open Kafelnikow Stich Tennis	1.0
Stich Tennis Viertelfinale	Achtelfinale Agassi Andre Agassi Australian Open Becker Boris Becker Courier Daviscup Doppel Edberg Einzel Endspiel Finale French Open gespielt gewann Graf Haas Halbfinale Huber Ivanisevic Kafelnikow Karbacher Match Melbourne Michael Stich Muster Nummer Pete Sampras Rothenbaum Runde Sampras Steeb Stefan Edberg Turnier Turniers US Open Weltrangliste Wimbledon Wimbledonsieger	0.6779

In einem scheinbar gleichen Gebiet kann es aber auch unterschiedliche feinere Cluster geben, was allerdings nicht immer der Fall sein muss:

w_1, w_2, w_3	$v_{1,2,3}$	C_v
Aminosäuren Blut Gehirn	Eiweiß Enzym Forscher Hormon Hormone Körper menschlichen Organismus Protein Proteine Substanzen Wissenschaftler Zellen	0.7406
genetischen Mutationen Virus	DNA DNS Erbgut Erbsubstanz Forscher Gen Gene Vermehrung Viren Wissenschaftler	0.8065
Krankheiten Pflanzen Wissenschaftler	Ausbreitung Bakterien bestimmte entwickeln Erbanlagen Erbgut erforschen Erforschung Erreger Forscher Gen Gene Genen genetisch genetische genetischen Gentechnik Gesundheit Insekten menschliche menschlichen Organismen Parasiten resistent Substanzen Tiere Tieren untersucht verschiedene verschiedener Viren zahlreiche Zellen übertragen	0.3887
Entstehung Galaxien Weltall	Astronomen Erde Kosmos Materie Planeten Sterne Universums Urknall	1.0
Atome Protonen Temperatur	Eigenschaften Elektronen Energie Forscher Masse Materie Physiker Strahlung Wasserstoff Wissenschaftler	0.5007
Atomkerne Elektron Physiker	Elementarteilchen Energie Protonen Quarks Teilchen	1.0
Außenminister Clinton Israel	Abkommen Albright amerikanische angekündigt Arafat Assad aufgefordert Barak bekräftigte Besuch besucht Beziehungen Bill Clinton Botschaft Christopher Einigung Erklärung Fortschritte Frieden Friedensprozeß Gespräche Gesprächen Haltung Irak Iraks Iran israelischen Israels Jassir Arafat Konflikt König Hussein kündigte Madeleine Albright militärischen Ministerpräsidenten Mubarak Nahen Osten Nahost-Friedensprozeß Netanjahu [..]	0.3163
Bekämpfung Frieden Israels	Abkommen Konferenz Terrorismus	1.0
Frieden internationale Netanjahu	Abkommen Außenminister Israel Nahen Osten Oslo Syrien Terrorismus Truppen US-Präsident Wahlen Washington	0.5886

Abkommen Armee tschetschenische	Boris Jelzin Flüchtlinge Grosny Jelzin Kaukasus-Republik Kontrolle Lebed Nachrichtenagentur Rebellen Republik russische Rußlands Streitkräfte Truppen Tschetschenen Tschetschenien Tschetscheniens tschetschenischen Unabhängigkeit unterdessen Waffen	0.6699
Kaukasus Rebellen russische	Afghanistan Alexander Lebed Armee Boris Jelzin Dagestan Gebiet Grosny Inguschetien Jelzin Jelzins Krieges Kämpfe Lebed Moskauer Putin Republik Russen russischer Streitkräfte Süden Truppen Tschetschenen Tschetschenien Tschetscheniens tschetschenische tschetschenischen Vorgehen	0.5551
Grosny Truppen tschetschenischen	Abkommen abtrünnigen Abzug Agentur Anatolij Kulikow angegriffen Angriff Angriffe Angriffen Argun Armee Artillerie Aslan Maschadow Bamut Berichten Berufung beschossen blutigen Boris Jelzin Dagestan Dorf Dschochar Dudajew Dudajews Dörfer eingekesselten eingenommen Einheiten Einnahme erbitterten erschossen Erstürmung Feuerpause fortgesetzt Freischärler Friedensgespräche Gefechte Gefechten geliefert Generaloberst Generalstabschef getötet Gratschow Grosnys Großoffensive Gudermes Hauptquartier heftigen Inguschetien [...]	0.2462

Es folgen einige weitere Beispiele für Worttripel, für die ein Mensch nicht-leere charakteristische Vektoren erwarten würde, welche aber keine oder nur wenige Wörter enthalten:

w_1, w_2, w_3	$\upsilon_{1,2,3}$	C_υ
Schiff Sojaöl Transport	-	0.0
Katastrophe Rohöl Tonnen	-	0.0
Meldung Polizei Sachschaden	-	0.0
Einbruch Schaden Versicherung	Diebstahl	0.0

Mit den erweiterten charakteristischen Vektoren lassen sich für die gleichen Tripel nicht-leere Ergebnismengen berechnen, doch wie erwartet, ist der Kontext nicht immer der mit dem Tripel unmittelbar referenzierte, da der Suchraum vergrößert wurde:

w_1, w_2, w_3	$v_{1,2,3}$	C_v
Schiff Sojaöl Transport	Bord brachten Fleisch gelagert Güter Hilfsgüter Hilfsgütern importieren Kilogramm Kleidung Konvoi lagern Lastwagen Lebensmittel Lebensmitteln Mais Medikamente Medikamenten Mehl Mengen Milchpulver radioaktiv Sojabohnen Srebrenica Stoffe transportieren transportiert Treibstoff täglich verarbeitet WFP Öl	0.1626
Katastrophe Rohöl Tonnen	-	0.0
Meldung Polizei Sachschaden	Brand Feuer gelöscht löschen	1.0
Einbruch Schaden Versicherung	Angeklagte aufgeklärt Autos Bad Cannstatt begangen bemerkt beschädigt bestraft Beute Charlottenhof Delikt Delikte Diebstahl Einbruch entwendet Erpressung ertappt Fahrzeug Fahrzeuge Fahrzeugen Feuerwehr fuhr Fälle Fällen Geldbörse gelegt gestohlen gestohlenen Göppingen Kaufhaus Kl. Kriminalität kriminelle kwa Körperverletzung Lastwagen lsw mittag Mord morgen nachmittag Nötigung Polizeiangaben Raub Sachbeschädigung Schaden Schecks schützen Straftaten Tat tausend Totschlag Täter,Dieb unverletzt Vaihingen Verbrechen Vergewaltigung vermutlich versuchte verurteilt,Polizeibericht vorgetäuscht Wagen Weilimdorf Wohnung	0.1754

6. Disambiguierung

Nachdem im vorangegangenen Kapitel charakteristische Vektoren von Worttripeln als deren inhaltlicher Kontext gedeutet wurden, ergibt sich daraus die Möglichkeit, einzelne Wörter in ihre jeweils einzelnen Bedeutungskomponenten oder in diesem Fall genauer – Kontextunterschiede zu zerlegen. Diese Aufgabenstellung kommt dem linguistischen Begriff des Disambiguierens eines Wortes nahe, ist aber nicht damit identisch.

Beim linguistischen Disambiguieren ist gemeint, dass wenn ein Wort mehrdeutig ist, dass dann die verschiedenen Bedeutungen oder entsprechende Kontexte aufgezählt werden, bzw. die passende Bedeutung in einem Satz gefunden wird. In dem vorliegenden Fall werden nicht echte verschiedene Bedeutungen aufgezählt, sondern verschiedene Kontexte, in denen dieses Wort benutzt wird. Dabei kann bei mehreren verschiedenen Kontexten immer wieder die gleiche Bedeutung des Wortes zugrunde liegen, muss aber nicht. Mit z. Bsp. ‘Schiff’ gibt es mindestens zwei Bedeutungen und zwar die einer bestimmten Dachkonstruktion in der Kirche und die des Fahrzeugs auf dem Wasser.

Mit der weiter unten beschriebenen Methode allerdings werden unter anderem zwei Kontexte errechnet, wobei das eine das Schiff zu Wasser und andere Fahrzeuge bezeichnet, während das andere die Güter, die im allgemeinen mit Schiffen transportiert werden, bezeichnet. Es ist also keine Disambiguierung im linguistischen Sinne, wird hier aber dennoch so benannt, weil es intuitiv ein passender Begriff ist und weil die echte Disambiguierung ebenfalls in den Ergebnissen beobachtet wird.

6.1. Überblick über den Algorithmus

Beim Disambiguieren in diesem Falle existiert als Eingabe genau ein Wort und als Ausgabe wird eine Menge von Wortmengen erwartet,

wobei die einzelnen Wortmengen Wörter enthalten, die jeweils zusammen untereinander und mit dem Eingabewort einen homogenen Kontext ergeben. An dieser Stelle werden folgende Eigenschaften der zugrundeliegenden Tripelmethode des vorigen Kapitels genutzt:

Zu drei Eingabewörtern findet es (bei Herausfilterung der long-range Wörter) entweder

- eine inhaltlich zusammenpassende Menge von Wörtern
- in einigen Ausnahmen (wenn alle drei Eingabewörter zusammen ambig waren) eine inhaltlich gemischte Menge von Wörtern oder
- nichts

Weiterhin werden die Satzkollokationen zu dem eingegebenen Wort genutzt, von denen bekannt ist, dass sie eine inhaltlich gemischte Menge von Wörtern darstellen, von denen die meisten einzeln mit dem Wort zusammen einen inhaltlichen Kontext bilden.

Es wird nun aus den Satzkollokationswörtern zusammen mit dem Eingabewort eine Menge von Worttripeln generiert. Zu jedem dieser Tripel kann nun ein charakteristischer Vektor berechnet werden. Aus den oben genannten Eigenschaften folgt, dass fast immer diese Vektoren Wörter homogenen Inhalts beinhalten. Die charakteristischen Vektoren werden dann nach Ähnlichkeit gruppiert und aus jeder Gruppe wird ein Vektor mit dem koordinatenweisen Maximum gebildet und das Ergebnis ausgegeben, wobei alle charakteristischen Vektoren sich jeweils nach ihrer Bedeutung gruppieren müssten, da sie oft große Mengen sich überschneidender Wörter haben.

6.2. Entfernung zwischen Wörtern und Clustern

Genauer handelt es sich hier um das Finden der sich in der Nähe eines Wortes befindlichen Cluster bzw. des Teilgraphen, bei welchem der clustering Koeffizient C_v am größten möglichen ist, bei möglichst vielen Wörtern (weil drei miteinander verbundene Knoten natürlich bereits maximale $C_v = 1$ besitzen, während die Wahrscheinlichkeit,

100 Wörter mit einem großen clustering Koeffizienten zu finden recht klein ist).

Def.: Entfernung $d(w_i, M_L)$ eines Wortes $w_i \in W_L$ zu einem Cluster M_L in dem Graphen Ω_L ist eine natürliche Zahl und bezeichnet die Anzahl von Verbindungen, die mindestens traversiert werden müssten, um von dem Wort mindestens ein beliebiges Wort $w_k \in M_L$ zu erreichen. $d(w_i, M_L) = 0$ bedeutet demnach, dass $w_i \in M_L$ ist.

Def.: Disambiguierung eines Wortes $w_i \in W_L$ in einem Graphen Ω_L ist eine Abbildung $\Lambda : w_i \rightarrow \{\kappa_1, \kappa_2 ..., \kappa_n\}$, die aus einem einzelnen Wort $w_i \in W_L$ in eine Menge von Kontextvektoren $\{\kappa_1, \kappa_2 ..., \kappa_n\}$ abbildet.

Da angenommen wurde, dass es sich bei dem Satzkollokationsgraphen um einen Graphen mit der small-world Eigenschaft handelt und die mittlere Weglänge in diesem Graphen ca. 5 Schritte beträgt, würde eine Suche nach Clustern in der Nähe eines Wortes w_i, wobei Entfernung $d(w_i, M_L) \leq 5$ ist, mit großer Wahrscheinlichkeit den gesamten Graphen liefern. Oder mit anderen Worten, innerhalb von etwa 5 Schritten ist nahezu jedes Cluster von einem beliebigen Wort erreichbar. An dieser Stelle wird der Fall betrachtet, zu einem Wort nach Clustern mit $d(w_i, M_L) = 0$ oder $d(w_i, M_L) \leq 1$ zu suchen.

Um auf jeden Fall exakt $d(w_i, M_L) = 0$ zu erreichen, wird verlangt, dass in allen generierten Tripeln eines der drei Wörter das Eingabewort ist. Durch die Art, wie der Tripelalgorithmus in seiner Trivialdefinition funktioniert, werden auch alle möglicherweise gelieferten Wörter direkt mit dem Eingabewort verbunden sein, was eine schärfere Form von $d(w_i, M_L) = 0$ bedeutet.

6.3. Generierung der Tripelmengen

Def.: Die Bildung der zu dem Eingabewort $w_i \in W_L$ relevanten Tripelmengen ist eine Abbildung $T : \Gamma_i \rightarrow \{\tau_1, \tau_2 ..., \tau_p\}$ mit $\tau_m = \{w_j, w_k, w_l\}$

und $w_j, w_k, w_l \in W_L$ aus der Menge der Nachbarn von w_i in eine Menge von Tripeln.

Die Spezifikation dieser Abbildung hat direkten Einfluss darauf, wie groß die maximale Entfernung der gefundenen Cluster sein wird:

- Es sei, um $d(w_i, M_L) = 0$ zu erreichen T folgendermaßen aufgebaut : Es werden mit $\Gamma_i = \{w_n, w_{n+1}, ..., w_{n+m}\}$ die unmittelbaren Nachbarn des Wortes w_i genommen und daraus maximal $\binom{|\Gamma_i|}{2}$ Paare[i] gebildet. Jedes Paar wird dann zum Tripel um das Eingabewort ergänzt. Es entsteht somit die folgende Menge:

$$\{\tau_1, \tau_2 ..., \tau_p\} = \{\{w_i, w_n, w_{n+1}\}, \{w_i, w_n, w_{n+2}\}, \{w_i, w_{n+1}, w_{n+2}\}, ..., \{w_i, w_{n+m-1}, w_{n+m}\}\}$$

- Für $d(w_i, M_L) \leq 1$: Es werden wieder mit $\Gamma_i = \{w_n, w_{n+1}, ..., w_{n+m}\}$ die unmittelbaren Nachbarn des Eingabewortes w_i genommen und daraus maximal $\binom{|\Gamma_i|}{3}$ Tripel gebildet, mit

$$\{\tau_1, \tau_2 ..., \tau_p\} = \{\{w_n, w_{n+1}, w_{n+2}\}, \{w_n, w_{n+1}, w_{n+3}\}, ..., \{w_{n+m-2}, w_{n+m-1}, w_{n+m}\}\}$$

Es lassen sich an dieser Stelle beliebige Funktionen zur Generierung von Tripeln definieren – die beste Mögliche für ein vorgegebenes festes $d(w_i, M_L)$ ist diejenige, die alle möglichen Kombinationen betrachtet. Allerdings kann die darauf folgende Berechnung der charakteristischen Vektoren dann zu aufwendig sein und daher kann es sinnvoll sein, Heuristiken zu suchen, die evtl. erkennen können, welche Tripel von Anfang an ignoriert werden können (siehe weiter unten). Interessant wären auch solche Definitionen , die Tripel „in die Tiefe“ statt in die Breite generieren, also 1. Wort ist direkter

[i] alle Kombinationen ohne Wiederholung

Nachbar von w_i, 2. Wort Nachbar vom 1. Wort und 3. Wort Nachbar vom 2. und dann die ersten beiden beibehaltend das letzte zu variieren. Das könnte dazu dienen, einen gefundenen Kontext genauer zu untersuchen, wurde allerdings im Rahmen dieser Arbeit nicht weiter verfolgt.

Weiterhin ist offen, welche Art von Kollokationen für die Bildung der Tripel genommen wird. Allein mit den vorhandenen Daten sind hier bereits mindestens 3 verschiedene Kollokationsarten möglich: linke Nachbarn, rechte Nachbarn und Satzkollokationen. Das Wählen einer anderen Kollokationsart kann an dieser Stelle auch prinzipiell andere Ergebnisse bringen. Allerdings handelte es sich dabei natürlich um einen anderen Graphen $\overline{\Omega}_L \neq \Omega_L$, der eventuell nicht kompatibel zu dem betrachteten ist und daher keine Rückschlüsse von $\overline{\Omega}_L$ nach Ω_L ermöglichen würde. Andererseits liegt beiden Graphen das gleiche Material zu Grunde – die beobachtete Sprache L, womit die Vermutung formuliert werden kann, dass Wissen über den einen auf dem anderen Graphen benutzt werden könnte. Dies stellt allerdings einen eigenen Forschungskomplex dar und konnte im Rahmen dieser Arbeit nicht durchgeführt werden.

Die Tripel selbst werden mit der in Kapitel 5.3 beschriebenen Approximation $X:\{w_i, w_j, w_k\} \rightarrow \upsilon_{ijk}$ in charakteristische Vektoren abgebildet. Diese Vektoren besitzen nun im Allgemeinen die Eigenschaft, „ähnlich“ zu sein, wenn sie aus dem gleichen Cluster stammen. Im vereinfachten Fall bedeutet dies, dass je mehr Wörter Vektor υ_1 und υ_2 gemeinsam haben, umso ähnlicher sind sie. Aufgrund dieser Ähnlichkeiten lassen sie sich gruppieren. Wenn also alle charakteristischen Vektoren υ nur jeweils Elemente aus dem ihnen zugrundeliegenden Cluster enthalten, sollte das Gruppieren aller charakteristischer Vektoren, die aus dem gleichen Cluster kommen, diese auch wieder zusammengruppieren.

Je mehr allerdings zwei Cluster $M_{L,1}$ und $M_{L,2}$ Elemente in ihrer Schnittmenge haben, umso schwieriger wird es, sie auseinander zu halten, weil ein charakteristischer Vektor, der aus einem solchen Tripel kommt, wessen sämtliche Elemente in der Schnittmenge liegen, auch Elemente aus beiden Clustern beinhalten wird[i]. Doch dadurch, dass die Schnittmenge mindestens 3 Elemente groß sein muss, wird die Wahrscheinlichkeit dafür in dem betrachteten Graphen bereits stark eingegrenzt.

Darüber hinaus werden große Mengen von Tripeln mit den dazugehörigen charakteristischen Vektoren generiert. Das führt dazu, dass wenn die Schnittmenge zwar vorhanden ist, aber nicht zu groß ist, dass dann die meisten Wörter entweder aus dem einen Cluster oder dem anderen kommen. Das Gruppierverfahren könnte diese Gegebenheit durch Gewichtung erkennen und zum Beispiel gemischte Vektoren identifizieren und ignorieren. In dieser Arbeit reichte allerdings bereits ein einfachstes Gruppierverfahren aus und in den seltenen Fällen von „Schnittmengentripeln“ wird der entsprechende charakteristische Vektor υ zu einer der beiden Seiten hinzugruppiert, was eine Verringerung der Precision für diese Fälle bedeutet.

6.4. Clusterverfahren

Im Allgemeinen werden Clusterverfahren verwendet, um eine Menge von Vektoren in disjunkte Gruppen zu zerlegen, mit dem Ziel, dass innerhalb der einzelnen Gruppen eine Homogenität oder Gleichheit existiert, aber zwischen den Gruppen ein wohldefinierter Unterschied zu sehen ist. Dazu gibt es eine Reihe unterschiedlicher Herangehenswiesen und diese lassen sich grob in hierarchische und nichthierarchische einteilen. Eine weitere grobe Einteilung aller Clusterverfahren wird durch den Charakter des Gruppierungsprozesses gegeben, wodurch zwischen aufbauenden (agglomerative) und abbau-

[i] siehe auch beispielhafte Disambiguierung von „Gold“ am Ende dieses Kapitels

enden (divisiven) unterschieden wird. In jedem Fall werden während des Prozesses die einzelnen Vektoren miteinander mit einem Distanzmaß verglichen. Eine gute Übersicht, sowie Definitionen finden sich in Läuter & Pincus [LP89].

An dieser Stelle wird ein hierarchisches agglomeratives Clusterverfahren (HAC) angewendet.

Bemerkung: Es wird im Folgenden $\vee$ als das koordinatenweise Maximum über zwei Vektoren benutzt.

Def.: Die Distanz *a* zweier Vektoren sei definiert als eine Funktion $\phi(\upsilon_1,\upsilon_2) \to 0 \le a \le 1$, welche zwei Vektoren υ_1 und υ_2 miteinander vergleicht und eine reelle Zahl *a* zwischen 0 und 1 liefert, wobei eine 1 bedeutet, dass die Vektoren identisch sind. Von den beiden Vektoren wird dabei zunächst der Kleinere[i] υ_k ausgewählt und mit dem anderen υ_g verglichen. Es wird die Anzahl der Einsen, die in beiden Vektoren übereinstimmen $\left|\upsilon_k \vee \upsilon_g\right|_1$ mit der Anzahl der Einsen $\left|\upsilon_k\right|_1$ des Kleineren verglichen:

$$a\left(\upsilon_k,\upsilon_g\right)=\frac{\left|\upsilon_k \vee \upsilon_g\right|_1}{\left|\upsilon_k\right|_1}$$

Dieses unsymmetrische Distanzmaß wurde aus dem Grund ausgewählt, dass es einfach und schnell zu berechnen ist. Es hat außerdem noch die Eigenschaft, dass wenn zwei Wortvektoren υ_1 und υ_2 verglichen werden und einer der beiden wesentlich mehr Wörter enthält, als der andere, also zum Beispiel $\left|\upsilon_1\right|_1 << \left|\upsilon_2\right|_1$ gilt, aber gleichzeitig nahezu alle Einsen, die in υ_1 enthalten sind, auch in υ_2 enthalten sind, also $\left|\upsilon_1 \vee \upsilon_2\right|_1 \approx \left|\upsilon_1\right|_1$, dass dann dieses Maß auch eine große Ähnlichkeit berechnet.

[i] nach Anzahl der Einsen gemessen

Mit Hilfe dieses Distanzmaßes lässt sich nun durch das Clusterverfahren die Menge der charakteristischen Vektoren $\{\upsilon_{ijk},...,\upsilon_{lmn}\}$ in einen Binärbaum überführen. Wenn man, einen Schwellwert x für die Distanz zugrundelegend, die Spitze des Baumes abschneidet, bekommt man eine Menge von Gruppen $\{G_1, G_2,..., G_n\}$ von charakteristischen Vektoren mit $G_i = \{\upsilon_1, \upsilon_2,..., \upsilon_j\}$. Aus jeder Gruppe wird nun ein Clustervektor κ_i folgendermaßen gebildet:

$$\kappa_i(G_i) = \upsilon_1 \vee \upsilon_2 \vee ... \vee \upsilon_j$$

Damit ist die Clusterabbildung aus einer Menge von charakteristischen Vektoren mit dem Distanzmaß a und dem Schwellwert x mit $\mathrm{K}_{a,x} : \{\upsilon_1, \upsilon_2 ..., \upsilon_J\} \rightarrow \{\kappa_1, \kappa_2,..., \kappa_n\}$ vollständig beschrieben.

6.5. Der vollständige Algorithmus

Die gesamte Disambiguationsabbildung Λ lässt sich nun noch einmal zur Übersicht kurz darstellen:

1. Eingabe eines Wortes $w_i \in W_L$.
2. Definieren der direkten Nachbarmenge Γ_i des Wortes w_i.
3. Generieren einer Tripelmenge $T : \Gamma_i \rightarrow \{\tau_1, \tau_2 ..., \tau_p\}$ aus der Nachbarmenge Γ_i.
4. Berechnung der charakteristischen Vektoren für alle Elemente von T, also Abbildung in ihre charakteristischen Vektoren: $\{\tau_1, \tau_2 ..., \tau_p\} \rightarrow \{\upsilon_1, \upsilon_2,..., \upsilon_n\}$.
5. Clustern der charakteristischen Vektoren mit dem Distanzmaß a und Schwellwert x und das Bilden des koordinatenweisen Maximums der Vektoren in den Gruppen, um die Clustervektoren zu erhalten: $\{\upsilon_1, \upsilon_2,..., \upsilon_n\} \rightarrow_{a,x} \{\kappa_1, \kappa_2,..., \kappa_m\}$.

Damit öffnen sich folgende Variierungsmöglichkeiten dieses Algorithmus:

- Bei 2. kann ein anderer Graph gewählt werden, um die Nachbarmenge zu definieren.
- Bei 3. kann eine Vorauswahl getroffen werden, welche Tripel zu nehmen sind und welche nicht.
- Bei 4. können verschiedene Methoden verwendet werden, um die charakteristischen Vektoren zu berechnen, siehe voriges Kapitel. Darüber hinaus könnte an dieser Stelle wieder ein anderer Graph verwendet werden.
- Bei 5. können verschiedene Distanzmaße verwendet werden, um einen besseres clustering Koeffizienten zu erreichen. Weiterhin kann mit Hilfe des Schwellwertes entschieden werden, ab wann der Clusterbaum abgeschnitten wird. Und schließlich kann nach jedem Hinzufügen eines Vektors zu einer Gruppe geprüft werden, wie sich der clustering Koeffizient C_v der Wörter der in einem Vektor abgebildeten Gruppe verhält. Wenn zwei echte eigenständige Cluster zusammengelegt werden, wird ein signifikanter Sprung des clustering Koeffizienten nach unten erwartet, da nahezu keine Verbindungen zwischen den Clustern gegeben sind.

Ferner gibt es bei dem Satzkollokationsgraphen eine Gewichtung der Verbindungen, die bislang zur Vereinfachung ignoriert wurde, doch ist es sinnvoll, sie nun wieder in die Diskussion mit aufzunehmen. Zum einen kann die Menge der Nachbarn Γ_i eines Wortes w_i recht groß werden, wobei dann die mit mindestens $\binom{|\Gamma_i|}{2}$ Kombinationen gebildete Tripelmenge entsprechend unvertretbar groß werden kann. In der Versuchsimplementierung dieses Algorithmus wurden aus diesem Grund die 1000 stärksten Verbindungen, also die mit dem größten Signifikanzwert, genommen.

Zum anderen kann diese Information an diversen anderen Stellen zur Verfeinerung genutzt werden, etwa bei der Bestimmung der charakteristischen Vektoren, oder beim Clustern.

Die Vielzahl der Kombinationsmöglichkeiten, die sich an dieser Stelle eröffnet haben, konnte im Rahmen dieser Arbeit leider nicht untersucht werden, nicht zuletzt auch deswegen, weil die Ergebnisse teilweise anderer Art sind und für die gegebene Aufgabenstellung nicht geeignet schienen. Doch für andere Aufgabenstellungen, sei es das Finden von synonymen Adjektiven, Kontexten auf Verbbasis u.s.w., gibt es hier noch viele Möglichkeiten.

6.6. Iteration

Schließlich ist wichtig, dass bei diesem Algorithmus diverse Teile iteriert werden können. Die Versuchsimplementierung beinhaltet eine solche Schleife:

Start mit $w_i \in W_L$ als Eingabe {

- Definieren der direkten Nachbarmenge Γ_i des Wortes w_i, wobei die 1000 stärksten Nachbarn genommen werden, es gilt somit $|\Gamma_i| \leq 1000$.

do {

- Generieren der Tripelmenge $T : \Gamma_i \rightarrow \{\tau_1, \tau_2 ..., \tau_p\}$ aus den stärksten 15 Elementen der Nachbarmenge Γ_i.
- Berechnung der charakteristischen Vektoren für alle Elemente von T, also Abbildung in ihren charakteristischen Vektoren: $\{\tau_1, \tau_2 ..., \tau_p\} \rightarrow \{\upsilon_1, \upsilon_2, ..., \upsilon_n\}$ mit solchen Tripeln, die das Eingabewort selbst wieder beinhalten.
- Clustern der charakteristischen Vektoren mit dem Distanzmaß a und Schwellwert 0.51 und Bilden des koordinatenweisen Maximums aus den Vektoren der

Gruppen, um die Clustervektoren zu erhalten: $\{\upsilon_1, \upsilon_2, ..., \upsilon_n\} \rightarrow_{a,x} \{\kappa_1, \kappa_2, ..., \kappa_m\}$. Außerdem wird noch in der Menge B gespeichert, welche Wörter zwar in Tripeln untersucht wurden, aber deren charakteristische Vektoren immer nichts enthielten.

- Entfernen aller in $\{\kappa_1, \kappa_2, ..., \kappa_m\}$ und in A vorkommenden Wörter und der gerade untersuchten 15 Elemente aus Γ_i.

} while $|\Gamma_i| > 0$

- Ausgabe von $\{\kappa_1, \kappa_2, ..., \kappa_m\}$.
- Ausgabe der Menge B.

} Ende.

Die Schleife in diesem Algorithmus lässt sich so begründen, dass wenn innerhalb einer betrachteten Menge von 15 Wörtern[i] mindestens zwei Wörter aus dem gleichen Cluster stammen, in dem auch das Eingabewort enthalten ist, also $w_i, w_j, w_k \in M_L$, wird dieses Paar zusammen mit dem Eingabewort einen charakteristischen Vektor mit den meisten Elementen aus diesem Cluster enthalten, je nach dem, wie stark dieses Cluster in der Tat zusammenhängt. Da diese Wörter aber ebenfalls allesamt Nachbarn des Eingabewortes sind, befinden sie sich auch in Γ_i und würden wieder untersucht werden, nur um wieder zum gleichen Cluster zu führen. In diesem Fall ist es also sinnvoll, diese gefundenen Wörter aus Γ_i zu entfernen, bevor sie untersucht wurden und damit den Rechenaufwand im günstigsten Falle deutlich zu verbessern, ohne an Precision zu verlieren.

[i] Oder eine beliebige andere Zahl - Situationsabhängig.

6.7. Ergebnisse

Hier werden einige beispielhafte Ergebnisse des Algorithmus vorgestellt und verschiedene Effekte besprochen. Betrachtet werden dabei stets alle die Clustervektoren, die mehr als vier Wörter beinhalten, denn es entstehen neben einigen Clustervektoren mit sehr vielen Elementen noch eine Reihe kleiner, mit meist einem bis vier Wörtern. Dabei sind die Clustervektoren stets nach der reinen Anzahl gefundener Wörter sortiert.

Zunächst wird das Beispiel angeführt, welches auch beim visualisierten Graphen eine deutliche Häufung für die hier gefundenen Clustervektoren aufweist:

Stich (nach 4. Durchlauf)

κ_1	ATP Achtelfinale Agassi Alleinspieler Andre Andrej Antwerpen As Asse Atout Aufschlag Australian Australier Ball Becker Beckers Bernd Biscayne Boris Break Breaks Brust Bälle Carl-Uwe Carlos Cedric Centre Chang Claus Coach Coeur Costa Courier Court DTB David Daviscup Daviscup-Team Daviscup-Teamchef Daviscup-Viertelfinale Doppel Doppelfehler Doppelpartner Dreekmann Duell Durchgang Edberg Einzel Elmshorn Endspiel Fehler Ferreira Finale Forget French Gegenspieler Gegner Goellner Goran Graf Grand-Slam-Turnier Grundlinie Guy Günter Haarhuis Haas Halbfinale Hendrik Henman [...] $C_{\kappa_1} = 0.2020$
κ_2	Alleinspieler As Bube Dame Fehler Gegenspieler Herz Herz- Herz-As Herz-Bube Herz-Buben Herz-Dame Herz-König Hinterhand K Karo Karo- Karo-As Karo-Bube Karo-Buben Karo-Dame Karo-Karten Karo-König Karte Kreuz- Kreuz-As Kreuz-Bube Kreuz-Buben Kreuz-Dame Kreuz-König König Mittelhand Pik Pik-As Pik-Bube Pik-Buben Pik-Dame Pik-König Siebter Stiche Trumpf Vorhand ausgespielte ausspielen bedient bekam fünften gestochen gewinnt gewonnen schmiert sechsten spielte stach sticht vierten wimmelt zieht zog übernimmt $C_{\kappa_2} = 0.2626$
κ_3	Bauch Brust Herz Messer Oberschenkel Schulter Stiche gestochen schlug stach verletzte versetzte zog $C_{\kappa_3} = 0.5172$
B	gelassenen fühlte gelassene ließen Groeneveld Herzgegend lasse Riglewski

Interessant erscheint in diesem Zusammenhang die Menge B, die außer zwei offenbar nicht zuordenbaren Namen und der ‘Herzgegend’ noch einige Formen der Redewendung ‘Im Stich gelassen zu werden’ enthält. Da diese Redewendung keine Bedeutung in dem Sinne hat, dass diese Redewendung stets in einem bestimmten Kontext genannt wird, bekommt sie an dieser Stelle auch keinen eigenen Clustervektor.

Dieser Fakt könnte in Zukunft genutzt werden, um echte Redewendungen automatisch erkennen zu können.

Neben solchen, durchaus positiv zu bewertenden Beispielen, gibt es auch Beispiele, die nicht so gut aussehen, an dieser Stelle sei das am Anfang des Kapitels erwähnte ‘Schiff’ aufgeführt:

Schiff (nach 14. Durchlauf)

κ_1	albanischen Anker Ankunft Atlantik Atommüll Atomtestgelände aufs ausgelaufen auslaufen Bahn Bay befand befördert Behörden beladen beladene Besatzung Besatzungsmitglieder Besatzungsmitgliedern beschlagnahmt Boot Booten Bord Braer Brand Brücke Bucht Bus Cherbourg Container Crew Dampfer Deck Decks Einfahrt eingelaufen eingetroffen Eisenbahn Elbe entladen ertranken Estonia Exxon Fähre fahren fahrende fahrenden Fahrt fährt Felsen Feuer Flagge Flotte Flüchtlinge Flugzeug Fracht Frachter Fregatte fuhr Fuß gebaut geborgen gefahren gekentert geladen gelaufen Genua gerettet geriet geschleppt Geschwindigkeit gesunken Gewässer Gewässern [...] $C_{\kappa_1} = 0.0838$
κ_2	Baumwollsaatöl Erdnußöl Fässern Leinsaatöl Mobile New Orleans Sojabohnenöl baldige fob loco prompte prompte-baldige raff roh unlesbar unverpackt $C_{\kappa_2} = 0.6885$
B	Propeller Dampfkraft zweites schlingernde Bordwand gestrandetes Knoten Mondsee Heimathafen Mesuf Überseehafen sinkendes untergegangen Monrepos-See vertäuen ankerte Hafenbehörde Rückfahrt Trog Badar Hafeneinfahrt flieht Amerika flott Passau kentert Blohm Rettungsboot hochseetüchtiges kapern Hansa-Hafen leck Fram Kap Hoorn trunkene schlingert Wasserlinie Reede gecharterte manövrierunfähig Holzmuschel angetrieben Hoi freigeschleppt Vega Liegeplatz Packeis ahoi Bergungsmannschaften besteigen Kemondo Endurance Dienst segelt schwimmende havarierte Sonargerät Eigner Grundsee Funkkontakt festmachen Trockendock enterten Frankonia Wellensiek Hal Över

Offenbar ist Schiff ein derart breitgefächerter Begriff, dass er in vielen recht verschiedenen Kontexten, die allerdings alle noch entfernt mit dem Wasserfahrzeug zu tun haben, genannt wird. Dagegen gibt es noch ein kleines Cluster mit einem hohen clustering Koeffizienten C_v, welches eine recht spezialisierte Bedeutung hat. Das Kirchenschiff dagegen konnte hier nicht gefunden werden, weil es wenig wahrscheinlich ist, dass in der Zeitung über ein solches geschrieben wird.

Man sieht, dass in der Menge B der Wörter, die gar nicht zugeordnet werden konnten auch „gute“ Wörter sind, nicht schlechter jedenfalls als die, die dem ersten Kontextvektor zugeordnet wurden. Offensichtlich haben sie aber keine passenden Querverbindungen zu ande-

ren Wörtern außer dem Eingabewort, wodurch sie ja erst hier auftraten, so dass sie nicht zugeordnet werden konnten.

Weiter es noch Wörter, die keine weiteren Kontexte, außer einem offensichtlichen besitzen:

Rauch (nach 14. Durchlauf)

κ_1	Alarm Anwohner Asche Brand Brandherd Brandort Bränden Dach Decke Dämpfe Explosion Explosionen Fenster Feuer Feuers Feuerwehr Feuerwehrleute Flammen Freie Gasen Gebäude Himmel Hitze Kamin Löschwasser Mitleidenschaft Qualm Ruß Schaden Staub Stock Treppenhaus Wind Wohnung Zigarette alarmiert ausgebreitet bemerkt bemerkte bemerkten beschädigt blies bläst brannte brennenden brennt entdeckt erstickt erstickten gemeldet giftigen plötzlich rasend $C_{\kappa_1} = 0.1914$
κ_2	Christopher Nivel Reifschläger Renger Tobias Zawacki $C_{\kappa_2} = 0.8653$
κ_3	Duft Geruch Himmel Nebel Qualm $C_{\kappa_3} = 0.1031$
B	100fach Albrights Aschenbecher Atemwege Bruggisser Dampf Del Feuchthaltemittel Feuermelder Feuern Freudenberg Fürstlicher Gebäudereste Glimmstengeln Glimmstengelprobleme Habemus Holzfeuer Hooligan-Szene Hooligans Häuserruinen Jugendzimmern Kalden Karoline Klodt Kommerzienrat Kontrollturm Kreosot Kulturamtsleiter Landwehrstr Levigion Lunge Lungenödem Lüftungsanlage Messingkanone Miami Möbelwerke München-Mitte Nebelmaschine Neo Notlandung Papstwahl Passagierraum Passivraucher Popocatepetl Pächterehepaar Rechtsmedizinerin Rosenau Räumen Saliger Salzsäure Schall Schaschlikspieße Schaum Schiiten-Dörfern Schlot Schloten Schornsteine Schornsteinen Schwelbrand Selbstgedrehten Stiegenhaus Teufelsdreck Theatertitanen Thürnau Toxische Tränengas Urbs Weißer Wronski Zeitungsfoto Zigarren Zugpausen ZÄ abgesaugt abziehen aufgehen aufgelöst aufgestiegen aufsteige aufsteigenden aufstieg ausgeatmeten ausstößt beißendem chemische dichte dichtem einatmen eingehüllt entweiche fumum geblasen gedrungen geht's gelber gerochen hineingeblasen inhaliert nichtkommerziell nikotinarm nikotinfrei nikotinfreies papam pusten qualmt saugte stieg verflogen verzieht waberte zischt

Hier sieht man bei κ_2 auch deutlich, dass Eigennamen oft eine besondere Rolle spielen. Es gibt eine Gruppe von Menschen, die in einem Zusammenhang in den Zeitungen auffällig oft gemeinsam genannt werden. Das trifft vor allem für Fußballer, Tennisspieler und andere Personen aus zeitungsrelevanten Gebieten zu.

In dem vorigen Kapitel wurde die Problematik der Schnittmengen zwischen Clustern am Beispiel von ‘Gold’, ‘Silber’ und ‘Bronze’ angesprochen:

Gold (nach 4. Durchlauf)

κ_1	100Meter 200Meter Abfahrt Albertville Almsick Aluminium Atlanta ausgezeichnet Australien Barcelona Barren Blei Bronze Bronzemedaillen Chinesin cif Dagmar Doppel dreimal Dreimal Edelmetall Edelmetallen Einzel Eisen Elfenbein EM Equipe Erdöl erfolgreichste erkämpfte errang europ Europameisterin Europameisterschaften Finale Franziska Freistil fünfmal g geholt gewann gewannen gewonnen Göteborg Gramm GUS Gwen Hase holte holten Kanada Keramik kg Kilogramm km Kombination kostete Kugelstoßen Kupfer Legierung Lillehammer loco m Mannschaftswettbewerb Marmor Medaille Medaillen Mehrkampf Messing Metall Nagano Nationenwertung neunmal Nickel notierte Öl Olympia Olympiade Olympiasieg Olympiasieger olympische olympischen Olympischen olympisches Paralympics Reißen Rennen Rettungsschwimmerabzeichen Riesenslalom Russin Schlußtag Schmuck schwamm Schweden Schweiz sechsmal Sekunden Seoul Sevilla sicherte siebenmal Sieger Silber Silbermedaille Silbermedaillen [Slalom Spielen Sprint Sprintstaffel Staffel Stahl Stoßen Südkorea Super-G Sydney Team Tonne Torrence Unze v.Chr. van Verkauf verkaufen Vierer viermal Wasserwacht Weitsprung Weltmeister Weltmeisterin Weltmeisterschaft Weltmeisterschaften Weltmeistertitel Weltrekord Weltrekordzeit] Winterspielen WM WM-Titel x Zink Zinn zweimal Zweimal $C_{\kappa_1} = 0.1500$
κ_2	abgebaut Aluminium Ankauf Antimon Blei Bodenschätze Bodenschätzen Bronze Chrom Degussa Diamanten Edelmetall Edelmetalle Edelmetallen Edelmetallpreise Edelsteine Eisen Eisenerz Erdöl Erze Erzen fein g gewonnen Gramm Gußeisen Kalkstein kg Kilogramm Kobalt Kohle Kupfer Legierung Mangan Marmor Mengen Messing Metall Metalle Metallen Mineralien Molybdän Münzen Nickel Palladium Perlen Platin Quecksilber reich Salz Schmuck Schwefel Silber Tonnen Unze Uran v.Chr. Vanadium Verkauf Vorkommen Wolfram x xx xxx Zink Zinn $C_{\kappa_2} = 0.2958$
κ_3	Edelmetall erhielt Feingold Feinunze fixiert Fixing gefixt kostete Londoner Nachmittags nachmittags Nachmittagsfixing Nickel notierte stieg Sydney Unze US-Dollar Verkauf verteuerte Vormittag Vormittags Vormittagsfixing Vortag Währung $C_{\kappa_3} = 0.3143$
B	18karätigem Blau Coast Feinunzen Fields geraubtes Helvetier Kt Midas Morgenstund Myrrhe purem Weihrauch Weißes

Offensichtlich wurde der aus dem ambigen Tripel ‘Gold’, ‘Silber’ und ‘Bronze’ entstandene charakteristische Vektor dem ersten Kontextvektor hinzugeclustert und hatte damit negative Auswirkung auf den clustering Koeffizienten. Es stellt sich nun die Frage, wie groß diese Auswirkung war. Wenn die falschen Wörter entfernt werden, steigt der Koeffizient auf $C_v = 0.1830$, was scheinbar kein großer Sprung ist, aber erstens ist der Koeffizient für die entfernte Menge selbst mit $C_v = 0.2699$ sehr groß, was für ein eigenständiges Cluster

spricht und zweitens ist das Entfernen von 26 zufälligen[i] Wörtern C_v = 0.1510 (also 0.68% Abweichung), dazu im Vergleich ist der Sprung mit der Entfernung der falschen Wörter mit 18.04% Abweichung doch spürbar. Eine empirische Untersuchung dieses Effekts ist allerdings kaum möglich, da in den Daten bisher zu wenige Beispiele für echt ambige Tripel wie dieses gefunden wurden.

Für ‘schrieb’ finden sich zwei Kontexte (κ_1 ist Schreiben in der Literatur und κ_2 ist Schreiben in der Presse), die ein Mensch eigentlich nicht trennen würde, die sich aber hier deutlich verschieden darstellen:

schrieb (nach 4. Durchlauf)

κ_1	Abhandlung Abhandlungen Artikel Arzt Aufsätze Auschwitz Ausgabe Autobiographie Autor Autorin Bestseller Biographie Brecht Briefe Briefen Bruder Buches Bände Bücher Büchern Bühnenstücke Chronik Conrad Deine Demokratie Dialoge Dichter Dichterin Dichters Drama Dramen Drehbücher Du Erinnerungen Erzählung Erzählungen Essay Essays Exil Fassung Feder Filme Filmen Filmmusik Freiheit Freund Freundin Friedrich Gedanken Gedicht Gedichtbände Gedichte Gedichten Gefängnis Geist George Geschichten Goethe Gott Hauptrolle Hauptwerk Henze Herausgeber Herz Hindemith Historiker Hitler Hofmannsthal Hugo Hörspiele Jahrhundert Jahrhunderts John Joseph Journalist Journalistin Juden Jugend Kapitel Kinderbücher Klavier Komponist Kompositionen Komödie Krieges Kritiker Kurzgeschichten König L La Le Lebens Leser Libretto Lieder Literatur Literaturkritik Lyrik Lyriker Maler Mann Mein Meine Memoiren Menschheit Mozart Nabokov Nation Nietzsche Novellen Oper [...] C_{κ_1} = 0.0830
κ_2	Artikel Ausgabe Autor Berufung Blatt Chefredakteur China Corriere Daily Der El Figaro Financial Gaseta Gazeta Gazzetta Guardian Haaretz Herausgeber Hürriyet Independent Irish Iswestija Journalist Journalistin Kolumne Kolumnist Kommentar Kommentator Kommersant Komsomolskaja Korrespondent L La Le Leitartikel Leser Londoner Magazin Mail Monde Mundo Nation Nesawissimaja New News Observer Pais Pariser Parisien Post Prawda Redakteur Repubblica Ribao Sabah Segodnja Sera Sport Stampa Standard Sun Tageszeitung Telegraph The Tiempo Times Verleger Washington Wochenzeitung Wyborcza York Yorker Zeitschrift Zeitung amtliche angesehene britische della englische englischsprachige erscheinende erschien französische konservative kürzlich las liberale linksliberale regierungsnahe römische spanische taz titelte veröffentlicht veröffentlichte veröffentlichten Überschrift C_{κ_2} = 0.1266
B	Autogramme Gästebuch Nachruf Theaterkritiken Zahlen rote schwarze

[i] Vereinfachend ein Bereich, der mit eckigen Klammern in der Tabelle markiert ist.

Außerdem gibt es hier wieder eine Redewendung: ‘schrieb rote/schwarze Zahlen’.

Zum Schluss noch ein Adjektiv, ‘grün’, um zu demonstrieren, dass sich Adjektive, ähnlich wie Verben, stets in einem breiten Gebrauchskontext finden, welcher daraus resultiert, dass sie eher eine Funktion als eine Bedeutung besitzen:

grün (nach 3. Durchlauf)

κ_1	Ampel Ampeln Bj Blau Blätter Exception FARBE Farbe Farben Fußgängerampel Gefieder Gelb Gesicht Glas Grün Haare Himmel Kyu Licht Merkmal Nationalfarben Oberseite Rot angestrichen blau braun bunt eingefärbt färben färbt gefärbt gefärbte gefärbten gelb gelbe gelblich grau km lackiert leuchten leuchtend leuchtet lila links orange pink rechts rosa rot rote roten rötlich schwarz schön verfärbt violett weißen $C_{\kappa_1} = 0.2537$
κ_2	Bäume Berge Gras Grün Himmel Hügel $C_{\kappa_2} = 0.5811$
B	Hexenrituale angehauchte angehauchten wählen

Hier wird der eher funktionale Gebrauch dieses Wortes deutlich, denn das erste gefundene Cluster beinhaltet Wörter, die zusammen nicht homogenen Inhalts sind, im Vergleich zu den bei Stich gefundenen Kontextvektoren.

Insgesamt lässt sich abschätzen, dass dieses Verfahren für Substantive immer mindestens einen sinnvollen Kontextvektor findet, außer das Substantiv ist zu selten. Allerdings sind die gefundenen Kontexte manchmal unerwartet oder nicht inhaltshomogen. Letzteres tritt nur bei den Frequenteren Substantiven auf. Die unerwarteten Kontexte hingegen lassen sich nicht verhindern. Sie rühren daher, dass der Benutzer andere Vorstellungen von der Bedeutung dieses Wortes hat, als wie dieses Wort in den hier zugrundeliegenden Zeitungen wirklich benutzt wird. Und in sehr seltenen Fällen, wie bei ‘Gold’, kann es passieren, dass zwar richtige und erwartete Kontexte gefunden werden, diese aber nicht sauber getrennt werden.

7. Vererbung von Sachgebieten

Mit einem Verfahren, welches die verschiedenen signifikanten Kontexte als Wortliste liefert, in welchen ein Wort gebraucht wird, lassen sich nun einige Herangehensweisen konstruieren, Sachgebieten, die bereits durch einige Wörter definiert sind, neue Wörter zuzuweisen. Dabei muss vor allem aber beachtet werden, dass das Verfahren nicht zwischen speziellen fachspezifischen Wörtern und allgemeingültigen Wörtern unterscheiden kann. Diese Unterscheidung müsste also, sofern gewünscht, noch nachträglich eingeführt werden.

Ob sie allerdings gewünscht ist, ist eine Frage, die nicht an dieser Stelle beantwortet werden soll. Sie hängt davon ab, ob einem Sachgebiet nun in der Tat lediglich für genau dieses Gebiet fachspezifische Wörter zugewiesen werden sollen oder ob eine für dieses Gebiet spezifische Wortliste aufgebaut werden soll.

Beispielsweise ginge es also um die Entscheidung, ob ‘Summe’ zu dem Sachgebiet ‘Versicherung’ gehören soll oder nicht. Zunächst muss festgestellt werden, dass dieses Wort ebenso gut in die ‘Mathematik’, in die ‘Ökonomie’ oder eine Vielzahl anderer Gebiete passt. Ist also eine sehr fachspezifische Wortliste gefragt, so hat ‘Summe’ zwischen ‘Rückversicherung’, ‘Haftpflichtversicherung’ und ‘Krankenversicherung’ natürlich nichts verloren. Anders verhält es sich, wenn Versicherungstexte mit Biologietexten aufgrund von Sachgebietszuordnungen von einzelnen Wörtern verglichen werden sollen (z. Bsp. für automatische Klassifizierung) – dabei kann eine Liste wie die bereits genannte zwar auch von Nutzen sein, wenn sie allerdings um ‘Summe’, ‘Prozent’ und ‘Schaden’ erweitert würde, könnten die Ergebnisse wesentlich verbessert werden.

Sehr spezielle Wörter von allgemeinen auf rein statistischer Ebene zu unterscheiden ist allerdings aufgrund der Zipfschen Gesetzmäßigkeit zufriedenstellend möglich. Es kann mit hinreichend großer Zuverlässigkeit davon ausgegangen werden, dass die Wörter mit der gering-

sten Frequenz aus einer gegebenen Wortliste auch die spezifischsten sind. In der Tat lässt sich beobachten, dass die in der Datenbank gespeicherten und weiter oben bereits genannten ‘*versicherung’ teilweise noch gar nicht, teilweise unter 5 mal gesehen wurden, wodurch sie auch im Kollokationsgraphen nicht auftreten. Anders wäre es natürlich, wenn die zugrundeliegenden Texte keine Zeitungen wären, sondern eine große Sammlung von versicherungsspezifischen Dokumenten (Gesetze, Richtsprüche, Arbeiten, Aufsätze, sogar Werbung und ähnliches). Wie sich weiter unten zeigen wird, ist es zusätzlich sinnvoll, zwischen den Wortarten zu unterscheiden – Adjektive, Verben und Funktionswörter lassen sich allgemein in zu vielen verschiedenen Kontexten benutzen und sind damit meist nicht zur Sachgebietszuordnung geeignet.

7.1. Eigennamen

Eine wichtige Fragestellung sind Eigennamen – eine „gewöhnliche“ Sachgebietseinteilung durch Menschen enthält bis auf wenige Ausnahmen keine Eigennamen. Ausnahmen davon sind z. Bsp.: ‘Aristoteles’, ‘Plato’ usw., bei denen mit dem Namen eine Bedeutung identifiziert wird, die bereits weit über die reine Referenz zu der jeweiligen Person oder dem jeweiligen Objekt hinausgeht. Im Allgemeinen wird angenommen, dass Namen nicht in Sachgebiete eingeteilt werden sollen, da sie austauschbar sind und in diesem Sinne keinen Inhalt tragen.

Es scheint aber im Gegensatz dazu, als würden Namen oft sehr spezifische Vertreter eines Sachgebietes zu sein, obwohl sie im Gegensatz zu sehr speziellen Begriffen wesentlich öfter beobachtet werden und damit noch besser geeignet sind, Sachgebiete voneinander abzugrenzen. ‘Helmut Kohl’ ist ein gutes Beispiel für einen sehr spezifischen Begriff für Politik, obwohl dessen Frequenz weit über den Frequenzen von politikspezifischen Begriffen liegt, wie etwa ‘Wahlbeteiligung’, ‘Parlamentswahlen’ u. a. Es handelt sich bei solchen Namen ebenfalls um solche, bei denen es nicht mehr um die reine Refe-

renz zu einem Objekt oder Wesen handelt, sondern um solche, bei denen mehr Bedeutung dahintersteht.

7.2. Passende Kontexte

Das Ziel, zu einem Sachgebiet passende Wörter zu finden, kann mit dem vorgestellten Disambiguierungsverfahren insofern erreicht werden, als dass bestimmt wird, welche Cluster in dem Kollokationsgraphen als zu diesem Sachgebiet passend gefunden werden können. Danach können nach Frequenz und Wortart passende Wörter aus diesen Mengen entnommen werden und dem gerade betrachteten Sachgebiet zugewiesen werden. Gegeben ist der Name des Sachgebietes, sowie einige Wörter, die bereits diesem Sachgebiet zugewiesen sind, und es gilt die zugewiesenen Wörter um neue zu erweitern.

Zu jedem Wort kann nun mit dem „Disambiguator“ eine Menge von Kontextvektoren $\{\kappa_1, \kappa_2, ..., \kappa_n\}$ gefunden werden, von denen erwartet wird, dass einer zu der Bedeutung dieses Sachgebiets passend ist. Hier muss also das Problem gelöst werden, zu entscheiden, welche von den gefundenen Vektoren für dieses Sachgebiet zutreffend sind und welche nicht. Dies wird erreicht, indem die Termvektoren der einzelnen Wörter untereinander wieder verglichen werden. Weiterhin muss die Möglichkeit bestehen, ein Wort ganz zu ignorieren, falls es in dem Graphen nicht mit der erwarteten Bedeutung assoziiert ist.

Wenn die Wörter aus dem gleichen Sachgebiet stammen, dann ist zu erwarten, dass sie sich auch in einigen gleichen Clustern befinden und daher müsste der Vergleich zweier Wörter erbringen, dass beide mindestens einen ähnlichen Vektor besitzen. Dieser wird dann jeweils ausgewählt. Mit anderen Worten – es wird gezählt, wie oft ein bestimmtes Cluster von diesem Sachgebiet erreicht wird und wenn eines mehr als eine bestimmte Anzahl mal referenziert wird, wird es genommen. An dieser Stelle sind offensichtlich Verfeinerungen möglich, es ist zum Beispiel ein Ranking denkbar, welches nicht nur

die Kontextvektoren der Wörter untereinander vergleicht, sondern auch die clustering Koeffizienten C_v, die Entfernungen der gefunden Cluster zu anderen Wörtern dieses Sachgebiets, usw. Wenn also ein Wort eine Menge von Kontextvektoren liefert, von welchen allerdings kein einziger ähnlich einem Kontextvektor von anderen Wörtern aus diesem Sachgebiet ist, kann es sein, dass dieses Wort gar nicht zu diesem Sachgebiet gehört.

Die Wörter aller solcherart gefundenen Termvektoren können danach als eine Gesamtmenge betrachtet werden, aus welcher die für dieses Sachgebiet zu nehmenden Wörter zum Beispiel per Frequenzfilter selektiert werden. Es kann nun geschehen, dass zu einem Sachgebiet ein Wort gefunden wird, wie auch für zwei andere Sachgebiete. An dieser Stelle muss entschieden werden, ob disjunkt zugewiesen werden soll, oder ob ein Wort in mehreren Sachgebieten entsprechend seiner Ambiguität auftreten darf. Das entspricht dem im zweiten Kapitel erläuterten Prinzip, dass der Algorithmus anpassbar sein sollte. Sofern allerdings ein Wort nur einmal zugewiesen werden soll, benötigt man eine Bewertung der Wörter – wie gut sie zu einem bestimmten Sachgebiet passen, damit das beste ausgewählt werden kann. Die Ergebnisse einer derartigen auf statistischem Wissen basierenden Bewertung kann allerdings oft überraschende und unerwünschte Ergebnisse bringen: ‘Scheibe’ etwa würde dann weder zur ‘Technik’ (als Unterlegscheibe), noch zum ‘Sport’ (als Wurfscheibe) hinzugerechnet werden, sondern zur ‘Astronomie’.

7.3. Sachgebietserweiterung

Def.: Definierende Wörter des Sachgebietes $s_i \in W_L$ ist ein Wortvektor ψ_i über W_L, bei welchem an all den Stellen i eine Eins steht, wenn das Wort w_i als zu dem Sachgebiet s_i zugehörig definiert wurde.

Als Sachgebiete werden wieder normale Wörter aus dem Graphen genommen. Die Definition kann so verstanden werden, dass eine sehr begrenzte Anzahl von Wörtern dieses Graphen als Zentren von

Sachgebieten definiert werden. Die definierenden Wörter für das jeweilige Zentrum, also das Sachgebiet, spannen dann den 'Bedeutungsraum' dieses Sachgebiets auf und es gilt alle die Cluster zu finden, die durch diese Aufspannung zu diesem Bedeutungsraum gehören.

Def.: Sachgebietserweiterung ist eine Abbildung $\psi_i \rightarrow \{(\kappa_1, v_1), (\kappa_2, v_2), ..., (\kappa_n, v_n)\}$ aus der Menge der definierenden Wörter eines Sachgebietes $w_j \in \Psi_i$ in eine Menge von Paaren (κ_n, v_n), bestehend aus einem Kontextvektor κ_n und einem Wortvektor v_n über W_L.

Bemerkung: Der Wortvektor v_n beinhaltet Information darüber, welches der definierenden Wörter in welchem Maße an dem Auftreten des Kontextvektors κ_n beteiligt war und zählt somit, wie oft dieses Cluster von dem Sachgebiet erreicht wurde.

Ein diese Abbildung realisierender Algorithmus gestaltet sich folgendermaßen:

Start mit $s_i \in W_L$ als Eingabe:

- ψ_i bezeichnet die Menge der bereits dem Sachgebiet s_i zugewiesenen Wörter.
- foreach $w_j \in \psi_i$ { Disambiguieren des Wortes $w_j \in W_L$ }
- Die nun entstandene Menge von Kontextvektoren $\{\{\bar{\kappa}_1, \bar{\kappa}_2, ..., \bar{\kappa}_m\}\}$ wird eineindeutig in eine Menge von Paaren abgebildet: $\{\{\bar{\kappa}_1, \bar{\kappa}_2, ..., \bar{\kappa}_m\}\} \rightarrow \{p_1, p_2, ..., p_n\}$ mit $p_1 = (\bar{\kappa}_1, \bar{v}_1)$, wobei $\bar{v}$ die Information darüber kodiert, aus welcher Menge von Kontextvektoren ein $\bar{\kappa}$ stammt. Das erste Element des Paares ist demnach stets ein Kontextvektor $\bar{\kappa}$. Das zweite Element ist ein Vektor über W_L. Dieser hat an der Stelle des Wortes, welches den Kontextvektor $\bar{\kappa}$

hervorgebracht hat, die Anzahl der in $\bar{\kappa}$ vorhandenen Einsen: $\bar{v} = (0,...,0,|\bar{\kappa}|,0,...,0)$.

- Diese Paare werden nun mit einem Clusterverfahren, die Kontextvektoren $\bar{\kappa}$ aus dem Paar p mit dem im vorigen Kapitel eingeführten Maß a vergleichend, gruppiert.
- Aus jeder Gruppe wird nun ein Vektorpaar erstellt, indem jeweils die Kontextvektoren $\bar{\kappa}$ aller Paare einer Gruppe mit der ebenfalls bereits erwähnten $\vee$ Operation zu einem neuen κ zusammengefügt werden und der Wortvektor aller Paare einer Gruppe $\bar{v}$ koordinatenweise zu einem neuen Wortvektor v addiert wird, es entsteht die folgende Menge von Paaren: $\{(\kappa_1, v_1),(\kappa_2, v_2),...,(\kappa_n, v_n)\}$.
- Ausgabe aller Paare, die in v an mehr als einer Stelle keine Null stehen haben.

Ende.

Das Laufzeitverhalten dieses Algorithmus lässt sich durch eine Beschränkung an die Durchläufe des Disambiguators stark beeinflussen. In den nachfolgenden Beispielen konnte diese Anzahl problemlos auf einen Durchlauf beschränkt werden. Der daraus erwachsende Nachteil, dass nur die unmittelbar stärkste Bedeutung eines Wortes gefunden werden kann, gleicht sich wieder damit aus, dass die Sachgebiete bereits größere Anzahlen (mehr als ein Dutzend) definierender Wörter besitzen.

Eine weitere Beschränkung ist die Filterung aller der Paare, die nur an einer Stelle bei v eine Zahl größer als Null besitzen. Wenn die Menge der das Sachgebiet s_i definierenden Wörter unrein ist und zum Beispiel ein Wort enthält, welches in seiner Bedeutung nicht in dieses Sachgebiet hineinpasst, dann wird es vermutlich auch nicht in seiner Bedeutung zu den anderen Wörtern passen und damit wird entsprechend keiner seiner Kontextvektoren κ zu einem der anderen passen.

Somit bleiben diese ungruppiert und behalten auch nur an einer Stelle eine Zahl ungleich Null in v.

7.4. Ergebnisse

Zunächst wird der Idealfall eines Sachgebietes angeführt, bei welchem in der Definitionsmenge eine ausreichende Anzahl von wirklich fachspezifischen Begriffen vertreten war und für die auch in dem Satzkollokationsgraphen ausreichend viele Cluster vorhanden waren.

Bei $\psi_{Medizin}$ sind die das Sachgebiet 'Medizin' definierenden Wörter eingetragen und die Wörter, die produktiv waren, welche also in ihrer Disambiguierung für dieses Sachgebiet relevante Cluster referenzierten, sind markiert.

Unter v findet sich jeweils eine Repräsentation des nach dem Gruppierungsprozess zusammengeführten v Vektors von Wörtern, die den unter κ angegebenen Vektor ergeben haben. Es wurden lediglich die vier ersten der nach $|\kappa|$ sortierten Paare $p=(\kappa,v)$ angezeigt, für welche gilt, dass in v mehr als an einer Stelle eine Zahl ungleich Null steht.

Medizin

$\psi_{Medizin}$	Chemotherapie Cholera Cholesterin Demenz Diabetes Diabetiker Diagnostik Dickdarmkrebs Diphtherie Embolie Endoskop Endoskopie Epidemiologie Epilepsie Erbkrankheit Eugenik Exitus Exzision Fettgewebe Fettsucht Fetus Fluor Früherkennung Fäkalien Gastritis Gehirnblutung Gehirnerschütterung Gelbfieber Gelbsucht Harnröhre Harnstoff Hepatitis Herpes Herzfrequenz Herzinsuffizienz Herzstillstand Hirnhautentzündung Hirnrinde Hypertonie Hämoglobin Immunschwäche Immunsystem Impfschutz Implantat Implantation implantieren Infarkt Infektion infizieren Infusion injizieren Inkontinenz Inkubationszeit intravenös Katheter Kleinhirn klinisch Klitoris Kollagen Koma Kontrastmittel Kontrollgruppe Kortex Kreislaufkollaps Kreißsaal Körpergewebe Lebensmittelvergiftung Leberkrebs Leberzirrhose Leukämie Luxation Lymphknoten Magersucht Mammographie medikamentös Medikation Medizin Melanom Meningitis Meniskusoperation Menopause Menstruation Mole Morbus Muskelschwund Nachsorge Neurose neurotisch Paranoia paranoid Pathologe Pädophilie querschnittsgelähmt Querschnittslähmung Schleimhaut Schuppenflechte Schädeldecke Selbstmedikation Sklerose Stammler Stenose

v	κ
Fettgewebe=4 Gelbsucht=1 Immunsystem=139 Implantation=1 implantieren=1 Infusion=2 injizieren=13 intravenös=2 Katheter=2 Körpergewebe=6 Lymphknoten=10 Melanom=8	Abwehr Abwehrkräfte Abwehrmoleküle Abwehrreaktion Abwehrsystem Abwehrzellen Aids Aidsvirus aktiviert angreift Antibiotika Antigen Antigene Antikörper Antikörpern Autoimmunkrankheit Bakterien befallen befallenen Behandlung bekämpfen bekämpft bestimmte bilden bildet Blut Blutkörperchen Blutzellen Diabetiker DNA Eindringlinge Eiweiß Eizellen Embryo Empfängers entwickeln Entzündungen Enzyme Erbgut erkannt erkennen erkranken Erkrankung Erkrankungen Ernährung Erreger Erregern Forscher Forschern fremde fremden gebildet Gehirn Gene genetisch gentechnisch geschwächtem geschwächten Gewebe gezielt Haut Hautkrebs hemmen HI-Viren HIV Immunabwehr Immunzellen Impfstoff Impfung Infektion Infektionen [...] C_{κ} = 0.1788
Herpes=2 Hirnhautentzündung=3 Infektion=97 infizieren=32 Inkubationszeit=15 medikamentös=6	Abwehrzellen Affen Aids Aids-Erreger Aids-Virus ansteckende Ansteckung Antibiotika Antikörper Antikörpern Arzt Atemwege ausbreiten Ausbreitung Ausbruch ausgelöst auslösen bakterielle bakteriellen Bakterien Bakterium behandeln behandelt Behandlung bekämpfen bekämpft beträgt Blut BSE Chlamydien chronische chronischen CJD entzündliche Erbrechen [...] C_{κ} = 0.1579
Cholesterin=3 Demenz=15 Diabetes=85 Diabetiker=7 Dickdarmkrebs=11 Epilepsie=17 Fettsucht=2 Gastritis=2 Hirnrinde=2 Hypertonie=4 Immunschwäche=6 Leberzirrhose=14 Magersucht=3 Morbus=7 Muskelschwund=1 Schädeldecke=1 Sklerose=33	Aids Allergien Alter Alzheimer Anorexia Arteriosklerose Arthritis Arthrose Aspirin Asthma Autoimmunkrankheiten Bauchspeicheldrüse behandeln Behandlung beta-1b Betaferon Betaseron Betroffenen Bewegungsmangel Blutdruck Bluthochdruck Cholesterin Cholesterinspiegel Cholesterinwerten chronisch chronische chronischen Demenzen Depressionen Diabetes Diabetikern Diagnose diagnostiziert Entstehung Epilepsie erhöhte erkranken erkrankt erkrankte erkrankten Erkrankung Erkrankungen Ernährung Faktoren Fettleibigkeit Fettstoffwechselstörungen Folgen Forscher Früherkennung Fällen Gehirn Gehirns Gicht Hepatitis Herz- Herz-Kreislauf-Erkrankungen Herzbeschwerden Herzerkrankungen Herzinfarkt Herzkrankheiten Herzleiden Hormon Hypertonie häufiger Immunsystem Immunsystems Infektion Infektionen Infektionskrankheiten Insulin Interferon Komplikationen Krankheit Krankheiten Krebs Leberzirrhose Leiden leiden leidenden leidet litt Malaria Medikament Medikamente MS Multiple Sklerose multiple Multiplen Multipler multipler Neurologe Osteoporose Parkinson [...] C_{κ} = 0.1798
Cholera=40 Diabetiker=2 Diphtherie=30 Gelbfieber=18 Gelbsucht=12 Hepatitis=29 Hirnhautentzündung=10 Impfschutz=10 Leberkrebs=1 Lymphknoten=4 Meningitis=19	Afrika Aids anderenorts Ansteckung Ausbreitung Ausbruch ausgerottet ausreichenden bakterielle Bakterien Cholera Dengue-Fieber Denguefieber Diphterie Diphtherie Durchfall Enzephalitis Epidemie Epidemien erkrankt erkrankte Erkrankungen Erreger Erwachsenen Fleckfieber FSME geimpft Gelbfieber Gelbsucht gestorben Haemophilus Hepatitis A Hepatitis HIV impfen Impfkommission Impflücken Impfschutz Impfstoffe Impfstoffen Impfung Impfungen Infektion Infektionskrankheiten infektiösen infiziert influenzae Keuchhusten Kinderlähmung klassifizierten Krankheiten Krebs Leber Leberentzündung Leberzirrhose leiden Lepra Lungenentzündung Malaria Masern Milzbrand Mumps näher Patienten Pest Pocken Polio Ruhr [...] C_{κ} = 0.1903

Bei dieser Operation entsteht noch ein Nebeneffekt, nämlich dass die neu gefundenen Wörter in verschiedene Gruppen aufgeteilt sind. Diese Aufteilung ist eine natürliche, dem Satzkollokationsgraphen immanente Aufteilung der Daten und kann von Nutzen sein. Eine sinnvolle Aufteilung des Sachgebiets 'Medizin' bieten sie aber bestenfalls für Journalisten, weniger für Mediziner. Da der Großteil der Daten hier allerdings auch Zeitungen entstammt, kann die Vermutung formuliert werden, dass eine spezielle medizinische Dokumentensammlung mit genügend großem Umfang an dieser Stelle eine sinnvollere Einteilung erbringen könnte. Ferner würde sie vermutlich auch Zusammenhänge wie zwischen 'Cholera', 'Diphterie' und 'Afrika' nicht oder nur im geringen Umfange beinhalten. Diese Vermutung konnte allerdings leider nicht im Umfang dieser Arbeit geprüft werden.

Ein Negativbeispiel wird an dieser Stelle ebenfalls angebracht. Es handelt sich hier um das Sachgebiet 'Autos'. Um die Auswirkungen schlechter Sachgebietsdefinitionsmengen besser zu verdeutlichen, wurden hier auch diejenigen Paare $p=(\kappa,v)$ zugelassen, bei denen v lediglich ein Wort enthielt:

Autos

ψ_{Autos}	Ambulanz Automobil Autos Borgia Bus Diesel Ente Gespann Jaguar Jeep Karosse Karre Karren Kombi Konvoi Kutsche Käfer Landauer Limousine Mini Minna Mobil Oldtimer Taxi Trabant Wagen Wohnmobil Wrack
v	κ
Konvoi=61	angegriffen Angriff Armee belagerte Belgrad Berg Igman beschossen Bihac Blauhelme bosnische bosnischen Bussen Djakovica Dorf eingetroffen Fahrzeuge Fahrzeugen Flüchtlinge Friedenstruppen gepanzerten getötet Gorazde Grenze Hilfsgüter Hilfsgütern hundert Jeeps jugoslawischen kroatischen Lastwagen Lebensmittel Lebensmitteln Lkw Marsch Medikamente Medikamenten Mehl Moslem-Enklave ostbosnischen Polizisten Prizren Rebellen Sarajevo Sarajewo Serben serbisch serbische serbischen Split Srebrenica südlich Tonnen traf Tuzla UN-Angaben UN-Flüchtlingshilfswerks UN-Soldaten UN-Sprecher UNHCR Zagreb Zepa C_{κ} = 0.2843
Mobil=41	Agip Amoco Aral beteiligt BP British Chevron Corp Corporation Dutch Esso Exxon Fusion Irving Joint-venture Konkurrenten

	Mineralölkonzerne Oel Oil Petrofina Petroleum Royal Shell Standard Stationen Tankstellen Tankstellengeschäft Texaco Total Treibstoffmärkten US-Konzern US-Ölkonzern Veba Wintershall Zusammenschluss Zusammenschluß zweitgrößte Öl Ölkonzern Ölkonzerne Übernahme C_K = 0.2770
Kombi=7 Limousine=30	3er A4 Allradantrieb Audi Avant Baujahr Bj BMW Cabrio Cabriolet Civic Coupé Ford Fünftürer fünftürige Kilogramm km km/h Kofferraum Kombi kW lieferbar Mercedes Modelljahr Motor Motoren PS Renault Sekunden T-Modell Wagen C_K = 0.2898
Käfer=24	Ameisen Arten Bienen Blattläuse Blüten Eier Falter Fliegen Heuschrecken Holz Hummeln Insekten Larven Motten Mundwerkzeuge Pflanzen Raupen Rinde Schmetterlinge Schnecken Spinnen Tiere Wespen Würmer C_K = 0.4815

Das einzige Paar, welches den Filter passiert hätte, hätte lediglich einige wenige neue Wörter gebracht. Alles andere hätte zurecht ignoriert werden müssen. ‘Käfer’ an dieser Stelle ist ein echt ambiges Wort, bei welchem aufgrund der Beschränkung auf nur einen Durchlauf des Disambiguator Verfahrens nur die erste und stärkste Bedeutung gefunden werden konnte. Da diese zu keinem anderen der das Sachgebiet definierenden Wörter passte, blieb es auch für sich, wie auch fast alle anderen Wörter dieses Sachgebiets.

Es lassen sich demnach folgende Bedingungen für ein „gutes“ Sachgebiet skizzieren:

- Die Menge der das Sachgebiet definierenden Wörter muss ausreichend groß sein.
- Der Disambiguator muss gegebenenfalls tief genug suchen dürfen – auf Kosten der Laufzeit.
- Die definierenden Wörter dürfen in ihrer Bedeutung nicht zu weit von einander entfernt sein.
- Verben und Adjektive müssten gesondert behandelt werden. Entweder wird es akzeptiert, dass ein Verb oder Adjektiv mehreren Sachgebieten gleichzeitig angehört (wie in den Beispielen gehandhabt), oder sie werden vollständig weggefiltert.

Die weiter oben eingeführte einfache Filterbedingung, dass in v mehr als an einer Stelle eine Eins stehen muss, reicht allerdings in manchen Situationen nicht aus und müsste nach einer empirischen Untersuchung der Daten erweitert werden:

Architektur

$\psi_{Architektur}$	Ablauf Anlauf Anschwellen Anzug Apsis Architektur Atlanten Atlas Aufriß Ausschmückung Balkontür Balustrade Baukörper Baumstamm Beletage bloß Brause Brüstung Bucht Bungalow Dachgarten Drum einteilen Empore Entlastung Erdmantel Erker Fassade Fertigungsauftrag Freitreppe Fries Front Fäden Galerie Gebälk Gebäude Genauigkeit Gewebe Glockenturm Grundstück Hals Haupt Helm kahl Kirchenschiff klassisch Knauf Knolle Krabbe Kreuz Kreuzgang Kulisse Kuppel Laterne Loggia Manierismus Maueröffnung Model Mosaikstein Nerv Nische Obelisk Oberlicht offen Ohr Ordnung Part Parzelle Patio Pavillon Penthouse Pergola Planung Plastik Plattform Plot Podium Portikus Relief Ring Rokoko Rosette Rücklage Saalbau Schiff Schlafstadt Schlußstein Springbrunnen Stadtbild Stadtplanung Stirn Strenge Terrassentür Türmchen Umgang Vorbau Vordach Vorhof weltlich Wintergarten
v	κ
Baukörper=1 Gebäude=277 Glockenturm=4 Kreuzgang=1 Kulisse=3 Kuppel=4 Schlafstadt=1 Stadtbild=16	abgerissen abreißen Abriß abzureißen Altstadt Anbau angemietet angezündet angrenzenden angrenzendes Anlage Anlagen Anschlag Anwesen Architekt Architekten Architektur Areal Areals aufgebaut Augenzeugen ausgebrannt Bauabschnitt Bauakademie bauen Bauherren bauliche baulichen Bausubstanz Bauten Bauwerk Bauzeit Bebauung Bebauungsplan Beben begonnen benachbarten Berlins beschossen beschädigt besetzt besetzten bestehenden Beton Bewohner Bibliothek Bombe Bomben Brand Brandstiftung brannte [...] C_{κ} = 0.0660
Galerie=227 kahl=1 Portikus=2 Oberlicht=2	10-18 11-19 12-18 14-18 Akademie Alten Rathaus Alten anläßlich Aquarelle Aquarellen Arbeiten art Arte Atelier Ateliers ausgestellt ausgestellten ausstellt ausstellte Ausstellungen Ausstellungseröffnung Ausstellungsraum Ausstellungsreihe Auswahl Barckhausstraße Baselitz besichtigen Betrachter Beuys Bietigheim-Bissingen Bildende Bildern Bildhauer Bildhauers Bleistiftzeichnungen blt Braubachstraße Buntentor Collagen Cézanne Dany Keller Di Di-Fr dienstags Direktor Do donnerstags [...] C_{κ} = 0.2940
...	...
Ring=1 Schiff=145	albanischen Anker Ankunft Atlantik ausgelaufen auslaufen Bahn Bay befand befördert Behörden beladen beladene Besatzung Besatzungsmitglieder Besatzungsmitgliedern beschlagnahmt Boot Booten Bord Braer Brand Brücke Bucht Bus Cherbourg Container Crew Dampfer Deck Decks Einfahrt eingelaufen eingetroffen Eisenbahn Elbe entladen ertranken Estonia Exxon Valdez fahren fahrende Fahrt Felsen Feuer Flagge Flotte Flugzeug Flüchtlinge Fracht Frachter Fregatte fuhr Fähre gebaut geborgen gekentert geladen [...] C_{κ} = 0.0984

Während das erste Paar noch akzeptabel war, sind es die anderen gezeigten nicht mehr, obwohl sie den einfachen Filter passieren würden. Für derartige Paare ist allerdings auffällig, dass es in *v* zwar mehr als eine Stelle mit einer Zahl ungleich Null gab, dass aber eine der Stellen deutlich größere Werte enthält, als die anderen. Aus dieser Beobachtung einen verfeinerten Filter zu erstellen, benötigt allerdings eine größer angelegte Untersuchung.

Allgemein lässt sich abschätzen, dass die meisten Sachgebiete in der Datenbank des Wortschatz-Lexikons zu wenige definierenden Wörter haben. Für die Sachgebiete allerdings, die genügend Wörter haben, funktioniert dieses Verfahren gut (wie bei ‘Medizin’), bis auf einige Ausnahmen. Diese Ausnahmen entstehen durch unpassende Wörter in der ursprünglichen Sachgebietsdefinition und führen in größeren Anzahlen dazu, dass sich die Bedeutung dieses Sachgebiets auf diese falschen Wörter verschiebt. Ein Beispiel ist die ‘Zoologie’, bei welcher sich in der Wortschatzdatenbank unter anderem folgende Wörter finden: ‘empfindlich’, ‘fein’, ‘Hämoglobin’, ‘Netzhaut’, ‘Neuaufbau’, ‘Perlmutt’, ‘Regeneration’ und ‘Zentralnervensystem’ und dazu führen, dass es zu einem Sachgebiet wird, welches identische Cluster referenziert, wie das Gebiet ‘Medizin’. Mit derartigen Daten kann eine vernünftige Behandlung dieses Sachgebietes nicht erfolgen.

8. Auswertung

Die in dieser Arbeit entworfenen Verfahren ermöglichen es, die verschiedenen Gebrauchskontexte eines Wortes in Form von inhaltlich homogenen Wortgruppen zu finden und Sachgebietszuweisungen einzelner Wörter auf andere zu vererben. Dadurch können große Mengen von natürlichsprachlichen Daten automatisch unter Umgehung des bisherigen Problems der Ambiguität der Wörter verarbeitet werden. Zusätzlich kann eine Einteilung großer Teile des Wortschatzes in Sachgebiete vor allem bei Klassifizierungssystemen verwendet werden.

Für optimale Ergebnisse müssen folgende Bedingungen erfüllt sein:

- Der Korpus der den Kollokationen zugrundeliegenden Texte muss entsprechend groß sein, dass statistische Beobachtungen für Kollokationen möglich werden.
- Der Inhalt der Texte sollte nicht zu sehr von dem anvisierten Einsatzgebiet abweichen, woraus folgt, dass für spezielle Anwendungsgebiete auch spezielle Datenquellen notwendig sind.

Wie auch das Zipfsche Gesetz der Beschränkung unterliegt, dass es für Wörter mit zu hoher oder vor allem zu niedriger Frequenz nicht ohne Weiteres gilt, so unterliegen die Ergebnisse der hier entworfenen Verfahren einer ähnlichen Beschränkung. Zu hochfrequente Wörter können problemlos in nahezu allen Kontexten benutzt werden und es ist daher nicht sinnvoll, diese zu suchen. Wörter, die in einem gegebenen Textkorpus zu selten vorkommen, können allgemein bei statistischen Verfahren unter Ausnutzung von Kollokationen nicht genutzt werden.

8.1. Ausblick

Es bieten sich hierbei vielfältige Anwendungsmöglichkeiten an. Eine aktuelle ist zum Beispiel die, nach sogenannten „Communities“ in In-

ternet zu suchen, was mit den in dieser Arbeit vorgestellten Verfahren zweifach zusammenhängt. Zum einen ist das Verfahren auf Graphen abstrahierbar, d.h. dass statt Knoten mit Wörtern gleichsetzen zu müssen, können für Knoten auch beliebige andere Objekte eingesetzt werden, wie etwa Webseiten. Damit eröffnet sich ein einfacher Algorithmus zum Finden von Clustern von Webseiten im Internet. Zum anderen liegt das Interesse bei Webseiten vor allem an dem Auswerten des beinhalteten Textes. Weitere Anwendungsmöglichkeiten umfassen automatische E-Mail Klassifikationen nach Sachgebieten, Verbesserung von Volltextsuchmaschinen, Verwaltung von Verkaufskatalogen, Vorbereitung von psycholinguistischen Experimenten und vieles mehr.

Schließlich kann die in dieser Arbeit verfolgte Methodik als eine weitere Annäherung der statistischen Methoden der Automatischen Sprachverarbeitung zu den traditionellen psycholinguistischen Modellen, wie überblicksweise bei Harley 1995 [Harley95] beschrieben, aufgefasst werden. Die Problematik der Wortspeicherung im Gehirn des Menschen wird bei einigen psycholinguistischen Modellen als eine graphenartige Struktur aufgefasst, zu den vor allem die konnektionistischen Modelle zählen. Mit dem in dieser Arbeit verfolgten Ansatz werden die betrachteten Daten ebenfalls als ein Graph betrachtet, in welchem die Wörter abgespeichert sind. Dabei lässt sich beobachten, dass die Art der Organisation der Wörter in diesem Graphen aus psycholinguistischer Sicht auf Zugriff optimiert ist: Von einem Wort können mit einem Schritt die zu diesem Wort inhaltlich passenden oder verwandten Wörter erreicht werden. Mit wenigen Schritten kann jedes beliebige andere Wortfeld erreicht werden. Diese Zusammenhänge zu untersuchen und für Psycholinguisten nutzbar zu machen, könnte Bestandteil weiterer Forschungen sein, ebenso wie eine Verfeinerung der konkreten Mechanismen.

9. Quellenverzeichnis

[Černý96] J. Černý (1996): Dějiny Lingvistiky [Geschichte der Linguistik]; VOTOBIA, Olomouc

[DDH90] S. Deerwester, S. T. Dumais, T. K. Landauer, G.W. Furnas, R. A. Harshman (1990): Indexing by latent semantic analysis; Journal of the Society for Information Science, 41(6), 391-407

[Devlin93] K. Devlin, Infos und Infone (1993): Die mathematische Struktur der Information; Birkhäuser Verlag

[Dietze94] J. Dietze (1994): Texterschliessung : Lexikalische Semantik und Wissensrepräsentation, K G Saur

[FCanSolé01] R. Ferrer i Cancho, R. V. Solé (2001): The Small-World of Human Language

http://www.santafe.edu/sfi/publications/

[Harley95] T. A. Harley (1995): The Psychology of Language; From Data to Theory; Psychology Press, Sussex

[Heyer95] G. Heyer, H. Haugeneder (1995): Language Engineering, Essays in Theory and Practice of Applied Natural Language Computing; Vieweg

[Lehr96] A. Lehr (1996): Kollokationen und maschinenlesbare Korpora : Ein operationales Analysemodell zum Aufbau lexikalischer Netze; Niemeyer, Tübingen

[LP89] H. Läuter, R. Pincus (1989): Mathematisch-statistische Datenanalyse; Akademie-Verlag Berlin

[Morton79] J. Morton (1979a): Word recognition. In J. Morton & J. C. Marshall Psycholinguistics series. Vol. 2: Structures and processes, London: Paul Elek, pp. 107-156

[Ruske94] G. Ruske (1993): Automatische Spracherkennung, Methoden der Klassifikation und Merkmalsextraktion; 2. Auflage, Oldenbourg

[Sanders96] M. Sanderson (1996): Word Sense Disambiguation and Information Retrieval; Proceedings of the 17th ACM SIGIR Conference, pp. 142-151

[Schmidt99] F. Schmidt (1999): Diplomarbeit, Automatische Ermittlung semantischer Zusammenhänge lexikalischer Einheiten und deren graphische Darstellung; Universität Leipzig

[SK64] P. Sgall a kolektiv (1964): Cesty moderní jazykovědy [Wege moderner Sprachwissenschaft]; ORBIS Praha

[TK87] M. Těšitelová a kolektiv (1987): *O češtině v číslech* [Über Tschechisch in Zahlen]; ACADEMIA Praha

[WStrog98] Watts, D.J., S.H. Strogatz (1998): Collective dynamics of 'small-world' networks; Nature 393:440-442

Erklärung

“Ich versichere, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.“

Leipzig, 10.06.2002 Stefan Bordag