Bordag Stefan

Vererbungsalgorithmen von semantischen Eigenschaften auf Assoziationsgrafen und deren Nutzung zur Klassifikation von natürlichsprachlichen Daten

Zusammenfassung: Das Ziel dieser Arbeit ist es, auf der Grundlage der im Projekt "Deutscher Wortschatz" generierten Assoziationsgrafen zwischen einer Wortform und ihren Satzkookkurenzen, ein allgemeines Verfahren zu entwickeln, mit welchem Wortformen disambiguiert und für eine automatische Sachgebietszuweisung genutzt werden können.

In diesem Aufsatz werden die wichtigsten Kernaussagen genannt und kurz erläutert. Für präzise formale Definitionen, ausführliche Diskussionen einzelner Variationen der Algorithmen sowie relevanter Problemstellungen sei auf die Diplomarbeit verwiesen.

1 Einleitung

Automatische Disambiguierung von lexikalisch ambigen Wörtern ist ein zweiseitiges Problem. Zunächst muss ein Wörterbuch erstellt werden, in welchem die zur Disambiguierung von Wortformen notwendige Information bereitgehalten wird. Hier sind demnach möglichst alle Bedeutungen einer bestimmten Wortform aufgelistet, zusammen mit einer Definition. Danach erst kann dieses Wörterbuch benutzt werden, um in einem Textabschnitt zu einer bestimmten Wortform die gerade benutzte Bedeutung zu bestimmen. Für diesen Teil der Disambiguierung existieren bereits diverse Algorithmen, siehe (Manning/Schütze 1999). Es stellt im Allgemeinen kein Problem mehr dar, die korrekte Bedeutung einer Wortform in einem Textabschnitt mit Hilfe eines statistischen Verfahrens zu bestimmen, sofern der Text genügend Anhaltspunkte liefert und die referenzierte Wortbedeutung für diese Wortform im Wörterbuch enthalten ist. Diese letzte Bedingung ist allerdings auch die am schwersten zu erfüllende, da das manuelle Erstellen eines wirklich vollständigen Wörterbuches sehr aufwändig ist. Daher würde ein Algorithmus durchaus von Nutzen sein und zum besseren Verständnis der statistischen Eigenschaften des Sprachgebrauchs beitragen, welcher dieses Wörterbuch zu einem gewissen Prozentsatz automatisch füllen könnte. Als ein Nebeneffekt würde ein derart vorgefülltes Wörterbuch wahrscheinlich beim Einsatz in Information Retrieval Systemen eine erhöhte Genauigkeit bringen, da hier die Definitionen eher dem entsprechen, wie die entsprechende Wortform benutzt wird.

Darüber hinaus ist es dann auch möglich, präziser mit Sachgebieten umzugehen. Da als Ergebnis des vorzustellenden Algorithmus eine Menge von Wortmengen entsteht, wobei jede wiederum für eine eigene Wortbedeutung der Wortform steht, kann aus der Zugehörigkeit einzelner Wörter einer solchen Menge auf die Zugehörigkeit des disambiguierten Wortes zu den Sachgebieten besser entschieden werden. Im Allgemeinen ist jedoch zu beachten, dass die Sachgebietsproblematik nicht trivial ist. Es gibt keine ideale Sachgebietseinteilung und es gibt ebenfalls keine generellen Regeln, wie eine gute Sachgebietseinteilung erstellt werden kann. Das Verfahren in dieser Arbeit geht von manuell erstellten Sachgebieten aus – läuft jedoch diese Einteilung der den Daten immanenten Einteilung zuwider, leidet die Güte der Ergebnisse.

Unter Ausnutzung früherer Arbeiten von Mitarbeitern an dem Projekt "Deutscher Wortschatz" über Eigenschaften von Kookkurrenzen von Wortformen in der Deutschen Sprache, siehe z. B. (Schmidt 1999) war es möglich, einen derartigen Algorithmus zu entwickeln, welcher ausschließlich auf statistischen Methoden basierend die offensichtlichsten Gebrauchskontexte einer Wortform trennen kann. Dies ist sogar mehr als verlangt war, denn zwei verschiedene Gebrauchskontexte können sich auf die gleiche Bedeutung beziehen, im Allgemeinen aber nicht umgekehrt.

Es gibt verschiedene Arten von Kookkurenzen von Wortformen. Eine wichtige Unterscheidung ist, wie weit links und rechts von einem Wort ausgehend die anderen Wörter in Betrachtung gezogen werden. In diesem Fall wurde immer der Satz als Grenze genommen. Eine weitere wichtige Unterscheidung ist, ob direkte linke oder rechte Nachbarn betrachtet werden, oder ob allgemeiner beobachtet wird, welche Wortformen auffällig oft mit dem betrachteten Wort auftreten. Bei dem in der Arbeit vorgestellten Algorithmus werden ausschließlich diese sogenannten Satzkookkurenzen benutzt, andere hätten ebenfalls benutzt werden können, doch die Eigenschaft der Symmetrie, die den Satzkookkurenzen im Vergleich zu den anderen zugrunde liegt, ist dabei aber wichtig.

2 Der grafentheoretische Ansatz

Der Algorithmus basiert wie erwähnt auf den genannten Satzkookkurenzen von Wortformen, sowie auf den kürzlich entdeckten Eigenschaften der sogenannten "small-world" Grafen, siehe (Watts/Strogatz 1998) für die erste Veröffentlichung zu diesem Thema und (Ferrero et al. 2001) für eine Anwendung ähnlich der hier vorgestellten. Ein Graf, welcher die Assoziationen zwischen den Wörtern einer Sprache darstellt, kann dadurch konstruiert werden, dass die Wortformen selbst als Knoten und Kookkurenzen von Wortformen als Verbindungen zwischen den entsprechenden Knoten betrachtet werden. Das heißt, dass zwei Knoten dieses Grafen verbunden sind, wenn die

zwei entsprechenden Wörter signifikant oft zusammen in Sätzen auftreten und nicht verbunden sonst. Der resultierende Graf hat scheinbar das Aussehen eines zufälligen Grafen. Es lässt sich allerdings ein wesentlicher Unterschied zu einem zufälligen Grafen zeigen, sowie im Gegensatz zum zufälligen Grafen auch zum regulären Grafen.

Dafür werden zwei Größen für alle drei Arten von Grafen berechnet und miteinander verglichen (die exakten Definitionen finden sich in Kapitel 4 der Diplomarbeit):

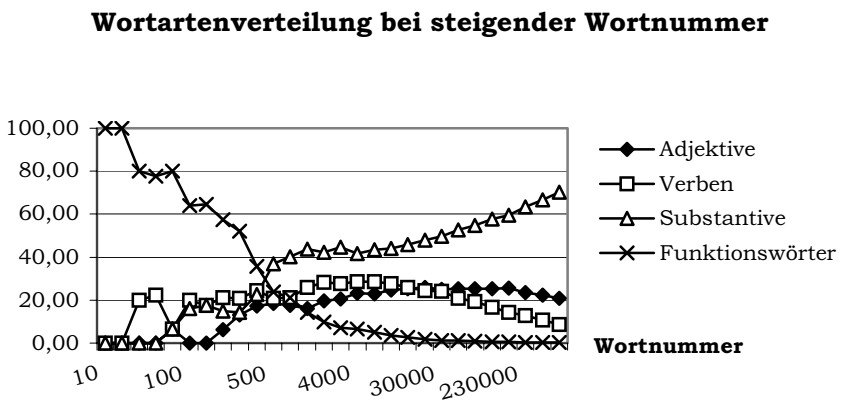
- Der clustering Koeffizient ist eine reelle Zahl und gibt über den gesamten Grafen gemittelt an, wie oft zwei unmittelbare Nachbarn eines Wortes auch untereinander wieder verbunden sind.
- Die mittlere Weglänge gibt an, wie viele Verbindungen mindestens zwischen einem beliebigen Wort und einem ebenfalls beliebigen anderen Wort liegen, gemittelt über den gesamten Grafen.

Bei einem zufälligen Grafen ist die mittlere Weglänge verhältnismäßig kurz und der clustering Koeffizient ebenfalls klein. Dafür ist bei einem regulären Grafen umgekehrt die mittlere Weglänge groß (da es keine Abkürzungen gibt) und dafür aber auch der clustering Koeffizient groß, da er an jedem beliebigen Punkt stärker gehäuft ist als der zufällige Graf an den meisten Punkten. Ein "small world" Graf ist in dem Sinne ein perfekter Graf, da er sowohl kurze Weglängen hat, als auch einen großen clustering Koeffizienten aufweist - bei gleich vielen Kanten und Knoten.

Eine besondere Anordnung ermöglicht es dem "small world" Grafen diese Eigenschaften zu haben. Diese besteht darin, dass der gesamte Graf aus ineinanderverschachtelten Clustern besteht, wobei ein Cluster eine Menge von Knoten ist, die auffällig mehr Verbindungen untereinander besitzt, als nach außen, zu anderen Knoten. Eine weitere wichtige Eigenschaft ist die, dass die sogenannten Stopwörter, zu welchen vor allem Artikel und häufig benutzte Verbformen gehören, die meisten Verbindungen besitzen (der Artikel ‚der' tritt z. Bsp. auffällig oft mit allen männlichen Substantiven auf) und dadurch mit jeder Gegend des Gesamtgrafen verbunden sind - sie stellen damit die Abkürzungen durch den Gesamten Grafen dar. Damit gestaltet sich ein typischer Weg von einem Knoten zu einem anderen (im Grafen) weit entfernten Knoten derart, dass ein Sprung zu dem Wort im gleichen Cluster folgt, welches mit einem Stopwort verbunden ist. Die Verbindung zum Stopwort liefert dann mit zwei weiteren Sprüngen (zum Stopwort und vom Stopwort weg) bereits einen Weg direkt in die Nähe des Zielwortes und mit einem weiteren Sprung wird bereits das Zielwort erreicht. Allein diese Erkenntnis könnte bereits interessant für die Psycholinguistik sein, welche sich unter anderem mit lexikalischen Zugriffsmethoden des Gehirns beschäftigt, da dies offensichtlich eine nahezu ideale Anordnung des Wortschatzes darstellt.

In der folgenden Abbildung sind die Wörter des Wortschatzes entsprechend dem Zipfschen Gesetz nach Häufigkeit des Auftretens auf der Abszisse abgetragen. Daraus wird deutlich, dass zu den häufigsten Wörtern die Funktionswörter der Sprache gehö-

ren, während die Inhaltstragenden Wörter seltener vorkommen, derer allerdings um Größenordnungen mehr existieren. In dem Grafen haben die Funktionswörter ebenfalls eine besondere Wirkung – die mittlere Weglänge bei wachsender Grafengröße nur sehr langsam mitwachsen zu lassen und im Gegenzug die lokale Gruppierungseigenschaft nur geringfügig einschränken.



2.1 Der Disambiguierungsalgorithmus

Die Anordnung der Wortformen in Cluster ausnutzend und unter Ausschluss der Funktionswörter, berechnet der Algorithmus die direkten Nachbarn einer Wortform, unterschieden in die verschiedenen beteiligten Cluster. Dadurch, dass die Cluster keine maximalen Cluster sind und dadurch, dass seltenst ein Wort mit einem anderen Wort immer zu den gleichen Clustern verbunden ist, wird für jedes Wort eine charakteristische Beschreibung jeder Wortbedeutung in Form von jeweils einer Wortmenge generiert. Es wird des Weiteren die Annahme formuliert, dass es vernachlässigbar wenige Worttripel gibt, für welche gilt, dass sie alle drei zusammen immer noch ambig sind. (Ein Beispiel ist ‚Gold Silber Kupfer' im Gegensatz zu ‚Gold Silber Bronze', wo bei dem zweiteren Tripel nach wie vor ambig bleibt, ob Olympische Auszeichnungen oder Metalle gemeint sind.)

Die Funktionsweise des Algorithmus lässt sich kurz und unformal folgendermaßen darstellen (die formale Definition findet sich im Kapitel 6.5 der Diplomarbeit):

1. Eingabe des zu disambiguierenden Wortes.
2. Bestimmen der direkten Nachbarn dieses Wortes.

3. Generieren einer Menge von Tripeln durch Kombination von Paaren der Nachbarn und jeweiligem Hinzufügen des Eingabewortes.
4. Bestimmen der Nachbarschnittmengen zu jedem Tripel. Dabei wird jedes Wort eines Tripels genommen, seine Kookkurenznachbarn bestimmt und über das Tripel deren Schnittmenge betrachtet.
5. Clustern aller Tripel nach ihren Nachbarschnittmengen.
6. Löschen der Nachbarschnittmengen.

An dieser Stelle können die Tripel innerhalb der Gruppen, in die sie aufgeteilt sind, aufgelöst werden und es entsteht eine Menge von Gruppen von Wörtern, wobei jede Gruppe einen eigenen Gebrauchskontext des Eingabewortes darstellt. Beim Clustern wurde ein Hierarchisches Agglomeratives Clusterverfahren (HAC) benutzt, siehe (Läuter/Pincus 1989) für einen Überblick über Clusterverfahren.

2.2 Beispiele und Diskussion der Resultate der Disambiguierung

Beispielhaft wurde die Wortform ‚Stich' disambiguiert, welche mehrere verschiedene Bedeutungen besitzt:

1. eine Verletzung
2. das Reizen beim Skatspielen
3. als Eigenname z. Bsp. 'Michael Stich'
4. der Kupferstich, wobei einzelne Werke als 'Stich' bezeichnet werden

Der Algorithmus war in der Lage, die ersten drei (und am meisten frequenten) der Bedeutungen zu finden und zu unterscheiden, sowie eine Gruppe von Wörtern B, welche keiner der drei Gruppen zugeordnet werden konnten. Darunter befinden sich auch die Kollokationen zu Stich aus der Redewendung 'im Stich gelassen zu werden'. Verallgemeinernd lässt sich beobachten, dass nahezu immer die Redewendungen in dieser nichtzuordenbaren Menge erscheinen werden, da die Redewendung meist mit der direkten Bedeutung der Wortform nichts zu tun hat, obwohl diese natürlich mit dem Wort zusammen auffällig oft auftreten.

Stich

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 [...]

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
κ_3	Bauch Brust Herz Messer Oberschenkel Schulter Stiche gestochen schlug stach verletzte versetzte zog
B	**gelassenen fühlte gelassene ließen** Groeneveld Herzgegend **lasse** Riglewski

Der Erfolg dieses Algorithmus hängt davon ab, auf welchem Korpus er operiert, im Verhältnis zu den erwarteten Ergebnissen. Der Korpus des Projekts "Deutscher Wortschatz" ist hauptsächlich aus journalistischen Medien aufgebaut, was in einer starken Überrepräsentierung gewisser Themenbereiche im Vergleich zu anderen resultiert, bis hin zu Bereichen die etwa gar nicht vorkommen. Dieser Fakt ist an dem Beispiel gut erkennbar, da für das in den Medien weit häufiger vorkommende Thema Tennisspielen und damit für die Bedeutung von Stich als Eigenname (‚Michael Stich') wesentlich mehr Wörter gefunden wurden als für die Bedeutung von Stich als Verletzung. Die Bedeutung des Stichs als Kupferstich war in dem Kollokationsgrafen überhaupt nicht auffindbar. Das zeigt auch deutlich, dass gewisse Themenbereiche, wie etwa Sport, Politik oder bis zu einem gewissen Maße auch Medizin mit diesem Korpus gut behandelt werden können, andere wieder, wie etwa Literatur, Kunst oder Wissenschaft, welche nur einen verschwindend geringen Teil der Medien einnehmen, nur schlecht behandelt werden können.

Insgesamt lässt sich abschätzen, dass dieses Verfahren für Substantive immer mindestens einen sinnvollen Kontextvektor findet, außer das Substantiv ist zu selten. Die gefundenen Kontexte sind allerdings 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 vorkommen, dass zwar richtige und erwartete Kontexte gefunden werden, diese aber nicht sauber getrennt werden, weil eine der weiter oben genannten Annahmen, mit denen der Algorithmus operiert, nicht immer zutreffen.

3 Sachgebietsvererbung

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 Kollokationsgrafen als zu diesem Sachgebiet passend gefunden werden können. Die Annahme an dieser Stelle lautet demnach wieder, dass die Cluster hinreichend oft inhaltshomogen sind. Danach können nach Frequenz und Wortart passende Wörter diesen Mengen entnommen werden und dem gerade betrachteten Sachgebiet zugewiesen werden.

Zu jeder Wortform kann nun mit dem "Disambiguator" eine Menge von Kontextwortmengen gefunden werden, von denen erwartet wird, dass mindestens eine zu der Bedeutung dieses Sachgebiets passend ist. Hier muss also das Problem gelöst werden, zu entscheiden, welche von den gefundenen Mengen für dieses Sachgebiet zutreffend sind und welche nicht. Dies wird erreicht, indem die Termvektoren der einzelnen Wortformen des Sachgebietes untereinander wieder verglichen werden. Weiterhin muss der Algorithmus die Möglichkeit haben, ein Wort ganz zu ignorieren, falls es in dem Grafen 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 akzeptiert.

Die Funktionsweise des Algorithmus für Sachgebietszuweisungen lässt sich kurz und unformal derart darstellen (die formale Definition findet sich im Kapitel 7.3 der Diplomarbeit):

1. Eingabe des zu erweiternden Sachgebietes.
2. Bestimmen der dem Sachgebiet bereits manuell zugewiesenen Wortformen.
3. Bestimmen der den einzelnen Wortformen zugehörigen Bedeutungswortmengen mit dem Disambiguierungsalgorithmus.
4. Die nun entstandene Menge von Bedeutungswortmengen wird eineindeutig in eine Menge von Paaren abgebildet, wobei ein Paar aus einem Wort und einer der dazugehörigen Bedeutungswortmengen besteht.
5. Diese Paare werden nun mit einem Clusterverfahren, wie beim Disambiguierungsalgorithmus bereits benutzt, gruppiert.
6. Aus jeder Gruppe wird nun ein Mengenpaar erstellt, indem alle ersten Elemente der Paare der Gruppe und alle zweiten Elemente der Paare jeweils zu einer Menge zusammengeführt werden.

Es entsteht demnach eine Menge von Mengenpaaren, wobei die erste Menge eines solchen Paares die Wortformen enthält, welche ein bestimmtes Cluster referenzieren und die zweite Menge die Wortformen, welche dieses Cluster darstellen.

3.1 Beispiele und Diskussion der Resultate der Sachgebietsvererbung

Ein interessanter Nebeneffekt dieses Verfahrens ist, dass bei der Erweiterung des Sachgebietes gleichzeitig eine natürliche, direkt den Daten entnommene Gruppierung der Wörter entsteht. Diese muss zwar nicht weiter beachtet werden, doch sie stellt eine Disambiguierung des Sachgebietes dar oder mit anderen Worten - potentielle Untersachgebiete werden damit gefunden. Als Beispiel wurde das Sachgebiet ‚Medizin' genommen.

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 der nach dem Gruppierungsprozess zusammengeführten Menge von Wörtern, die die unter κ angegebenen Menge von Gebrauchskontextwörtern ergeben haben. Es wurden lediglich die vier ersten der nach $|\kappa|$ sortierten Paare $p = (\kappa, v)$ angezeigt.

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** [...]
v	κ
Fettgewebe Gelbsucht Immunsystem Implantation implantieren Infusion	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 Blutzel-

injizieren intravenös Katheter Körpergewebe Lymphknoten Melanom	len 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 [...]
Herpes Hirnhautentzündung Infektion infizieren Inkubationszeit medikamentös	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 [...]
Cholesterin Demenz Diabetes Diabetiker Dickdarmkrebs Epilepsie Fettsucht Gastritis Hirnrinde Hypertonie Immunschwäche Leberzirrhose Magersucht Morbus Muskelschwund Schädeldecke Sklerose	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 [...]
Cholera Diabetiker Diphtherie Gelbfieber Gelbsucht Hepatitis Hirnhautentzündung Impfschutz Leberkrebs Lymphknoten Meningitis	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 [...]

4 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:

- Das 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.

4.1 Ausblick

Es bieten sich hierbei vielfältige Anwendungsmöglichkeiten an. Eine aktuelle ist zum Beispiel die, nach sogenannten „Communities“ im Internet zu suchen, was mit den in dieser Arbeit vorgestellten Verfahren zweifach zusammenhängt. Zum einen ist das Verfahren auf Grafen 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) be-

schrieben, aufgefasst werden. Die Problematik der Wortspeicherung im Gehirn des Menschen wird bei einigen psycholinguistischen Modellen als eine grafenartige Struktur aufgefasst, zu den vor allem die konnektionistischen Modelle zählen. Mit dem in dieser Arbeit verfolgten Ansatz werden die betrachteten Daten ebenfalls als ein Graf betrachtet, in welchem die Wörter abgespeichert sind. Dabei lässt sich beobachten, dass die Art der Organisation der Wörter in diesem Grafen 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.

Literatur

Ferrero, R. et al. (2001): The Small-World of Human Language. Online im Internet: http://www.santafe.edu/sfi/publications/.

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

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

Läuter, H./Pincus, R. (1989): Mathematisch-statistische Datenanalyse. Berlin: Akademie-Verlag.

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

Manning, C. D./Schütze, H. (1999) Foundations of statistical natural language processing. Cambridge, MA: MIT Press.

Ruske, G. (1993): Automatische Spracherkennung, Methoden der Klassifikation und Merkmalsextraktion. 2. Auflage, München, Wien: Oldenbourg.

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

Schmidt, F. (1999): Automatische Ermittlung semantischer Zusammenhänge lexikalischer Einheiten und deren grafische Darstellung, Diplomarbeit, Universität Leipzig.

Sgall, P. et al. (1964): Cesty moderní jazykovědy [Wege moderner Sprachwissenschaft]. Praha: ORBIS.

Těšitelová, M. Et al. (1987): *O češtině v číslech* [Über Tschechisch in Zahlen]. Praha: ACADEMIA.

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