Reinhard Rapp

Die Erkennung semantischer Mehrdeutigkeiten mittels Unabhängigkeitsanalyse

1 Einführung

Die größte Schwierigkeit bei der maschinellen Verarbeitung natürlicher Sprache besteht darin, dass – im Gegensatz zu Programmiersprachen – die meisten sprachlichen Elemente mehrdeutig sind. Dies trifft auf vielen Ebenen der Verarbeitung zu, beispielsweise der phonologischen, der orthografischen, der morphologischen, der syntaktischen und der semantischen Ebene. Es scheint nun, dass die erst in jüngster Zeit entwickelte *Unabhängigkeitsanalyse* (engl. Independent Component Analysis, ICA), einen Ansatzpunkt zur Lösung dieses Problems liefert (Hyvärinen et al. 2001; Roberts & Everson 2001). Die Unabhängigkeitsanalyse kann als Weiterentwicklung der bekannteren Hauptkomponentenanalyse im Hinblick auf die Berücksichtigung statistischer Abhängigkeiten höherer als zweiter Ordnung betrachtet werden.

Das prototypische Anwendungsgebiet der Unabhängigkeitsanalyse ist die Signalverarbeitung, in der sie zur Trennung von Signalquellen verwendet wird. Dies kann am Beispiel des *Cocktail-Party-Problems* demonstriert werden. Dabei geht es darum, ein Sprachsignal aus den Störgeräuschen einer lauten Umgebung herauszufiltern. Ähnliches leistet die Unabhängigkeitsanalyse. Sie ist in der Lage, zwei Signale *A* und *B* zu rekonstruieren, wenn zwei unterschiedliche Mischungen dieser Signale vorliegen. Dies geschieht durch die Berechnung der *unabhängigen Komponenten* der Mischsignale. Zur Trennung der Signale zweier Schallquellen ist es also lediglich erforderlich, zwei an verschieden Standorten positionierte Mikrofone aufzustellen. Entsprechend sind zur Trennung von drei Signalquellen drei Mikrofone erforderlich usw.

In der vorliegenden Arbeit soll am Beispiel der bislang nur selten angegangenen Induktion von Wortbedeutungen gezeigt werden, dass sich diese Vorgehensweise auch auf Probleme im Bereich der maschinellen Verarbeitung natürlicher Sprache übertragen lässt. Dabei besteht die Zielsetzung darin, die möglichen Bedeutungen eines mehrdeutigen Wortes maschinell zu bestimmen. Im Gegensatz dazu sind beim in der Literatur viel häufiger behandelten Disambiguierungsproblem die möglichen Bedeutungen bereits bekannt, und es geht lediglich darum, die im jeweiligen Kontext richtige Lesart zu ermitteln.

Die Gliederung der Arbeit ist wie folgt: Zunächst wird ein kurzer Überblick über das im Bereich der Computerlinguistik noch kaum bekannte Verfahren der Unabhängigkeitsanalyse gegeben. Dann werden einige mögliche Ansätze zur Induktion von Wortbedeutungen vorgestellt. Einer dieser Ansätze wird anschließend genauer besprochen und einige Ergebnisse werden vorgestellt. Zum Schluss werden die Vor- und Nachteile der Vorgehensweise diskutiert und es wird ein Ausblick gegeben.

2 Unabhängigkeitsanalyse

Die Unabhängigkeitsanalyse ist ein statistisches Verfahren, das es erlaubt, versteckte voneinander unabhängige Faktoren zu finden, wenn nur lineare Überlagerungen dieser Faktoren beobachtet werden können. Die für die Unabhängigkeitsanalyse verwendeten Algorithmen sind noch immer Gegenstand der Forschung, und für die genauere Darstellung der theoretischen Grundlagen existieren umfangreiche Werke (Hyvärinen et al. 2001; Roberts & Everson 2001). Wir beschränken uns deshalb hier darauf, die Wirkungsweise anhand einiger prägnanter Beispiele zu veranschaulichen.

Zunächst betrachten wir die beiden in Abb. 1 (a) gezeigten Signale, nämlich eine Sinus- und eine Sägezahnschwingung. Diese beiden Signale wurden, wie in Abb. 1 (b) gezeigt, durch eine gewichtete Addition gemischt, wobei das obere Mischsignal in diesem Fall als *1,5 s1 + s2* und das untere Mischsignal als *s1 + 2 s2* berechnet wurde (die Wahl der Faktoren ist unerheblich). Die beiden Mischsignale wurden anschließend in Form zweier Vektoren der Unabhängigkeitsanalyse zugeführt. Verwendet wurde der FastICA-Algorithmus nach Hyvärinen et al. (2001), auf den in Abschnitt 4.4 näher eingegangen wird. Zu beachten ist, dass in den Abbildungen zur Veranschaulichung zwar kontinuierliche Signalkurven dargestellt sind, dass die Signale jedoch tatsächlich nur jeweils durch 41 Zahlenwerte (Punkte in den Kurven) repräsentiert werden. Das heißt, die Eingabe für den Algorithmus besteht aus zwei Vektoren mit je 41 Einträgen. Weitere Informationen stehen dem Algorithmus nicht zur Verfügung.

Der Algorithmus versucht, die Mischsignale in der Weise zu trennen, dass sich für die getrennten Signale eine maximale Unabhängigkeit ergibt, d.h., der Verlauf des einen Signales soll keinerlei Rückschlüsse auf den Verlauf des anderen Signales zulassen. Dieser Versuch resultiert in den in Abb. 1 (c) dargestellten Signalen. Diese sind – bis auf die Multiplikation mit einem Faktor, der auch negativ sein kann – mit den (dem Programm unbekannten) Ausgangssignalen weitgehend identisch. Dies bedeutet, es ist gelungen, die Ausgangssignale aus den Mischsignalen zu rekonstruieren. Dies funktioniert in entsprechender Weise auch für mehr als zwei Signale. Es werden jedoch immer mindestens so viele Mischsignale benötigt, wie unabhängige Komponenten zu berechnen sind.

Abb. 2 zeigt jedoch, dass eine solche Rekonstruktion der Ausgangssignale nicht in allen Fällen möglich ist. Zwar handelt es sich bei den Ausgangssignalen wieder um eine Sinus- und eine Sägezahnschwingung, und auch die Mischsignale wurden auf dieselbe Weise erzeugt. Diesmal sind die Frequenzen der beiden Signale jedoch gleich. Dies bedeutet, dass die bei der Unabhängigkeitsanalyse geforderte Bedingung, dass die Signale zu jedem Zeitpunkt unabhängig voneinander sein müssen, schon bei den Ausgangssignalen nicht erfüllt ist. (Die starke Abhängigkeit ist daran ersichtlich,

dass jedem Signalpegel des Sägezahns ein eindeutiger Signalpegel der Sinusschwingung zugeordnet ist, wenn auch nicht umgekehrt). Dies führt dazu, dass die berechneten unabhängigen Komponenten in diesem Fall ihrem Namen nicht gerecht werden (d.h. nicht wirklich unabhängig sein können) und mit den Ausgangssignalen nicht identisch sind.

Obwohl der Gedanke nahe liegt, dass bei der Unabhängigkeitsanalyse die Kontinuierlichkeit des Signalverlaufes ausgenutzt werden könnte, so ist dies tatsächlich nicht der Fall. Abb. 3 (a) zeigt zwei mittels eines Zufallsgenerators erzeugte Signalverläufe, die also keinerlei Kontinuierlichkeit aufweisen, jedoch die Forderung nach Unabhängigkeit perfekt erfüllen. In Abb. 3 (c) ist zu sehen, dass diese Signale mittels Unabhängigkeitsanalyse gut rekonstruiert werden konnten. Es werden also keinerlei Bedingungen an den Signalverlauf gestellt. Wichtig ist lediglich die Unabhängigkeit der Signale zu jedem Zeitpunkt.

Abb. 4 schließlich gibt einen Hinweis, wie die Unabhängigkeitsanalyse funktioniert. Hier sollen Signale getrennt werden, die nur aus wenigen Impulsen bestehen, was auch gelingt. In Abb. 4 (b) ist zu sehen, dass die Mischsignale erkennen lassen, welche Impulse eine Variation in dieselbe Richtung aufweisen. Der dritte und der fünfte Impuls müssen deshalb zum selben Signal gehören, weil sie in den beiden Mischsignalen jeweils in dieselbe Richtung variieren (Kovarianz). Entsprechendes gilt für den ersten, den zweiten und den vierten Impuls. Werden nun diese Variationen schrittweise so lange vergrößert, bis sich ein maximaler Unterschied zwischen den beiden Signalverläufen ergibt, so erhält man die unabhängigen Komponenten.

Obwohl die in Abb. 1 gezeigte einwandfreie Signaltrennung recht erstaunlich ist, mag die Unabhängigkeitsanalyse nach den Betrachtungen von Abb. 4 fast schon einfach erscheinen. Da die Unabhängigkeitsanalyse jedoch nicht nur auf zeitabhängige Signalverläufe anwendbar ist, sondern sich für beliebige Vektoren eignet, sind ihre Einsatzmöglichkeiten fast unbegrenzt. Sie reichen von der Bildverarbeitung (Trennung überlagerter Bilder und Mustererkennung), über Messungen der elektrischen Aktivitäten des Gehirns (wo angebrachte Elektroden nur Mischsignale aufnehmen können) bis hin zur Wirtschaft (z.B. zur Erklärung der Verläufe von Börsenkursen).

Ganz allgemein kann gesagt werden, dass in der Natur beobachtete Ereignisse sehr oft Überlagerungen der Wirkungen mehrerer Ursachen sind. In der Vergangenheit bedurfte es immer wieder genialer Eingebungen, um diese Ursachen zu erkennen. Ein Beispiel aus der Physik ist etwa die Newton'sche Mechanik, deren Fallgesetz nur im Vakuum exakt beobachtbar ist, so dass der Einfluss der Luftreibung auf die Bewegung eines Körpers vom Einfluss der Schwerkraft getrennt werden muss. Ein Beispiel aus der Biologie sind die Mendel'schen Gesetze, die die Weitergabe (Überlagerung) von Erbanlagen beschreiben. Es ist zu erwarten, dass die Unabhängigkeitsanalyse einen Beitrag leisten kann, solche Überlagerungen zu erkennen.

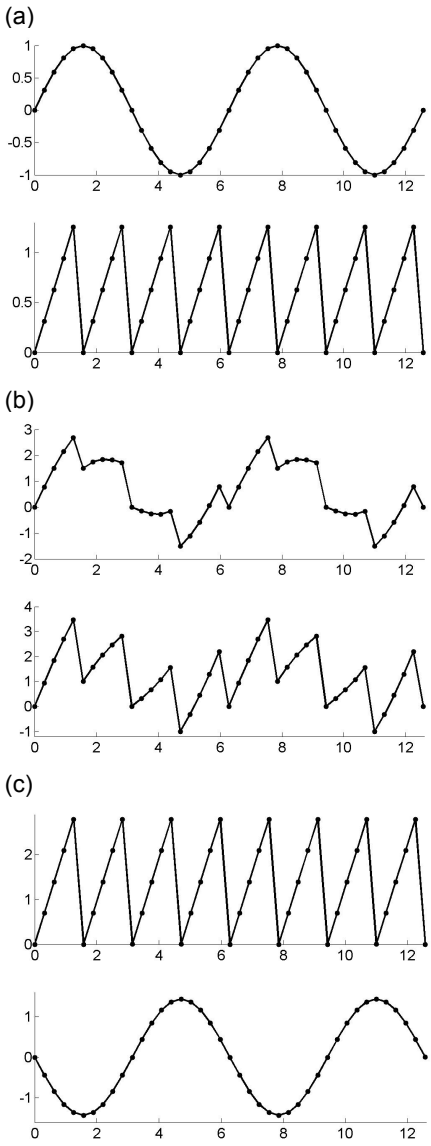

Abb. 1: Signaltrennung mittels ICA.

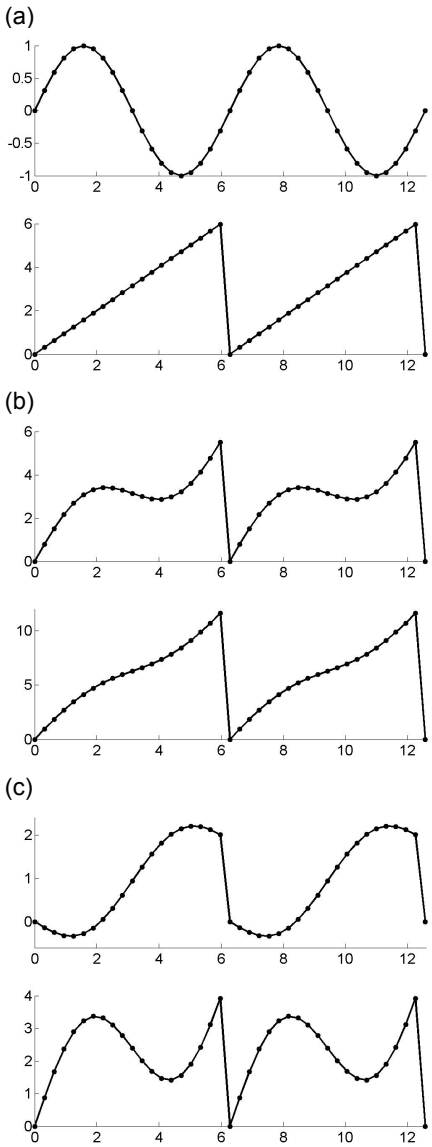

Abb. 2: Versuch mit abhängigen Signalen.

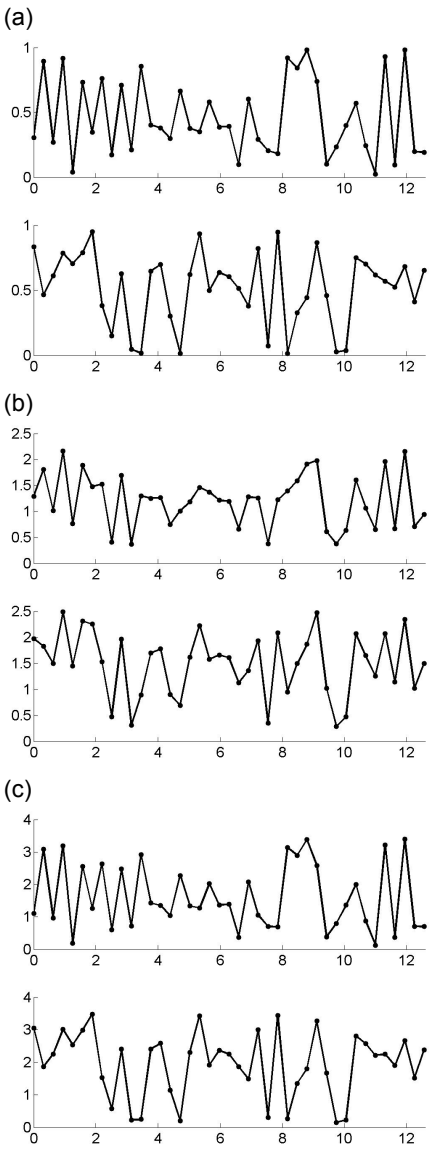

Abb. 3: Trennung zufälliger Signale.

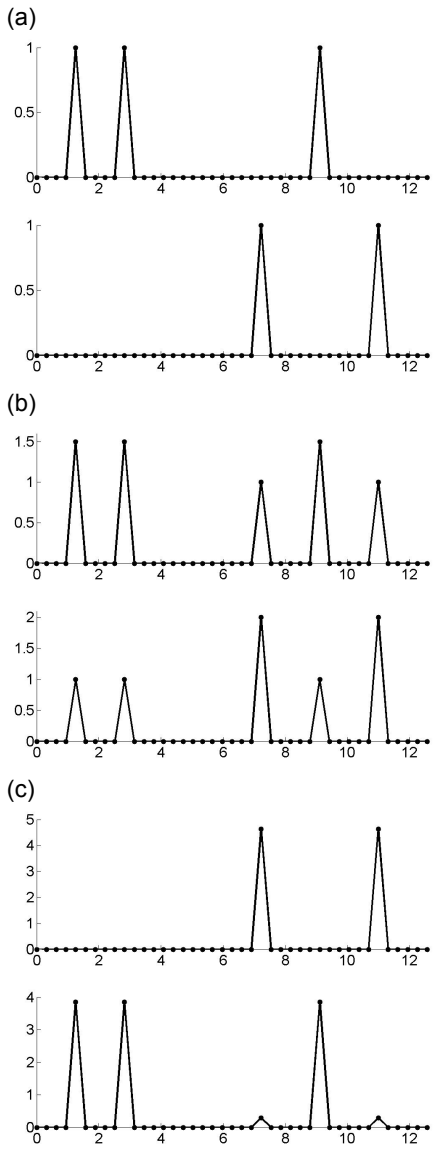

Abb. 4: Trennung impulsartiger Signale.

3 Ansätze zur Induktion von Wortbedeutungen

Ein Großteil der korpuslinguistischen Literatur zum Thema Wortbedeutungen beschäftigt sich mit der Disambiguierung mehrdeutiger Wortformen. Dabei wird davon ausgegangen, dass die möglichen Bedeutungen eines Wortes bereits bekannt sind und es darum geht, die im jeweiligen Kontext korrekte Lesart zu ermitteln. Dies gilt auch für Veröffentlichungen, in denen von „unsupervised word sense disambiguation" die Rede ist (z.B. Yarowsky 1995), denn der Begriff „unsupervised" meint in diesem Zusammenhang lediglich, dass zur Durchführung der Kontextanalyse kein manuell mit Wortbedeutungen annotiertes Beispielkorpus zur Verfügung steht.

Der erste uns bekannte Versuch, Wortbedeutungen zu induzieren, stammt von Arns (1994). Sie versuchte, die von Versuchspersonen genannten Assoziationen[1] auf mehrdeutige Wörter auf der Basis eines einfachen Assoziationsmaßes erster Ordnung zu clustern und hoffte, auf diese Weise Cluster zu erhalten, die mit den Wortbedeutungen korrespondierten. Dies gelang auch im Prinzip. Aus heutiger Sicht ist allerdings klar, dass mit anderen Assoziationsmaßen erster oder besser zweiter Ordnung, etwa dem log-Likelihood Ratio oder dem Cosinus-Koeffizienten, bessere Ergebnisse erzielt werden könnten.

Während Arns nur die typischen Assoziationen zu einem mehrdeutigen Wort clusterte, wurde in neuerer Zeit der Versuch unternommen, semantisch orientierte Cluster aller Wörter einer Sprache zu bilden (Lin & Pantel 2002). Soweit Clustering-Methoden verwendet werden, die es erlauben, dass ein Wort mehreren Clustern angehören kann, bedeutet dies ebenfalls das Induzieren von Wortbedeutungen. Jeder Cluster, dem das Wort angehört, wird dabei als eine seiner Bedeutungen interpretiert. Der Nachteil dieser Methode besteht – im Unterschied zu der von Arns – darin, dass nur eine eingeschränkte Anzahl von Bedeutungen möglich ist, die der Anzahl der Cluster entspricht.

Es ist aber eine Vielzahl weiterer Ansätze zur Induktion von Wortbedeutungen denkbar. Im Folgenden führen wir einige auf, die wir für vielversprechend halten, auf die wir in der Literatur aber noch nicht gestoßen sind. Um die Darstellung zu vereinfachen, wird teilweise nur auf den Fall Bezug genommen, dass ein mehrdeutiges Wort genau zwei Bedeutungen hat.[2] Generell wird davon ausgegangen, dass sich die jeweiligen Bedeutungen eines Wortes am besten durch möglichst charakteristische Kontext-

[1] Die Methode ist gleichermaßen auch auf künstlich berechnete Assoziationen anwendbar.

[2] Die Erweiterung auf eine feste Anzahl mehrerer Bedeutungen kann in naheliegender Weise geschehen. Schwieriger wird es allerdings, wenn die Anzahl der Bedeutungen nicht vorgegeben ist. In diesem Fall können die Algorithmen für jede denkbare Anzahl angewandt werden. Unter den erhaltenen Ergebnissen wird dasjenige ausgewählt, das einerseits möglichst viele Bedeutungen unterscheidet, bei dem diese aber nicht zu dicht beisammen liegen.

wörter ausdrücken lassen. Beispielsweise kann das Wort *Bank* einerseits im Sinne von *Geldinstitut*, andererseits im Sinne von *Sitzgelegenheit* interpretiert werden. Bedeutungstypische Kontextwörter sind in diesem Falle aber weniger die genannten Synonyme zu *Bank* (da das gemeinsame Auftreten eines Wortes mit einem Synonym z.B. im selben Satz eher selten ist), sondern beispielsweise Wörter wie *Geld* oder *Park*.

1. *Disjunkte Auftretenspositionen:* Die von Yarowsky (1995) prägnant als „one sense per document" formulierte Beobachtung, dass mehrdeutige Wörter im selben Dokument meist nur in einer Bedeutung vorkommen, kann wie folgt ausgenutzt werden: Zwei Wörter gelten dann als gute Deskriptoren der Bedeutungen eines zweideutigen Wortes, wenn jedes von Ihnen zwar überzufällig häufig mit diesem Wort gemeinsam auftritt, wenn sie aber dennoch selten zusammen vorkommen. Die Suche nach solchen Wortpaaren könnte durch Maximieren des folgenden Ausdrucks realisiert werden:
   $K(A, \neg B) * K(B, \neg A) / (K(A, B) * K(\neg B, \neg A))$
   Hierbei ist $K(A, \neg B)$ die Anzahl der Umgebungen (Dokumente, Sätze oder Fenster), in denen zwar Wort *A*, nicht aber Wort *B* vorkommt. Erste Versuche mit diesem Ansatz zeigten, dass diese Vorgehensweise zwar im Prinzip funktioniert, dass sie aber darunter leidet, dass nur ein sehr kleiner Teil der in den Korpora vorhandenen Kookkurrenzinformationen genutzt werden.
2. *Dokumentklassifikation:* Gegeben ein zweideutiges Wort, besteht das Ziel darin, die Dokumente eines Korpus, die dieses Wort enthalten, in der Weise in zwei Mengen aufzuteilen, dass die Überlappung der beiden Mengen bezüglich der signifikanten Assoziationen zu dem mehrdeutigen Wort so gering wie möglich ist. Für das Beispiel *Bank* enthielte dann die eine Menge Wörter, die mit der Bedeutung *Geldinstitut* zusammenhängen, die andere solche, die mit der Bedeutung *Sitzgelegenheit* zusammenhängen. Die beiden Bedeutungen des mehrdeutigen Wortes berechnen sich als die Positionen im semantischen Raum, die sich durch Überlagerung der Kookkurrenzvektoren seiner Assoziationen in der jeweiligen Dokumentmenge ergeben.
3. *Komplementäre Merkmale:* Ausgangspunkt der Überlegungen ist die Annahme, dass sich der Kookkurrenzvektor eines mehrdeutigen Wortes durch Addition der Kookkurrenzvektoren seiner Bedeutungen ergibt, was sich dadurch rechtfertigen lässt, dass die lexikalische Umgebung eines mehrdeutiges Wortes die Vereinigungsmenge der lexikalischen Umgebungen seiner Bedeutungen sein sollte. Das Problem bei der Induktion der Wortbedeutungen ist nun, dass zwar der Summenvektor, bekannt ist, dass aber nicht klar ist, welches die beiden Summanden sind. Eine Möglichkeit, die Summanden zu bestimmen, besteht darin, eine Vielzahl möglicher Wortpaare darauf hin zu untersuchen, inwieweit die Summe ihrer beiden Kookkurrenzvektoren mit dem Vektor des mehrdeutigen Wortes übereinstimmt. Eine optimale Übereinstimmung ist zu erwarten, wenn sich die Vektor-

einträge der beiden Summanden exakt ergänzen, d.h. wenn jeder der beiden Vektoren genau diejenigen Merkmale aufweist, die dem anderen fehlen. Das so gefundene Wortpaar eignet sich zur Beschreibung der beiden Bedeutungen des mehrdeutigen Wortes.

4. *Varianz der Kookkurrenzhäufigkeiten:* Wenn über mehrere Korpora (oder längere Dokumente) verschiedener Textsorten verglichen wird, welche anderen Wörter mit einem vorgegebenen mehrdeutigen Wort wie häufig gemeinsam auftreten, so wird bei Wörtern, die mit den unterschiedlichen Bedeutungen zusammenhängen, eine besonders hohe Varianz der Häufigkeiten zu beobachten sein. Beispielsweise wird in einem landwirtschaftlichen Mitteilungsblatt das Wort *Speicher* häufig im Zusammenhang mit *Silo* oder *Getreide*, hingegen aber selten zusammen mit *Computer* oder *Festplatte* auftreten, während es in einer Computerzeitschrift genau umgekehrt sein mag. Wörter, bei denen sich hohe Häufigkeitsunterschiede beobachten lassen, sind offenbar für die Wortbedeutungen charakteristisch. Um jedoch entscheiden zu können, welche dieser Wörter derselben, und welche unterschiedlichen Bedeutungen zuzuordnen sind, müssen noch die Schwankungsrichtungen einbezogen werden. *Computer* und *Festplatte* gehören deshalb zur selben Bedeutung von *Speicher*, weil in den Texten, in denen *Computer* häufig vorkommt, auch *Festplatte* häufig auftritt, und weil in jenen Texten, in denen *Computer* selten vorkommt, auch *Festplatte* selten ist.

Der im Rahmen dieser Arbeit vorgestellte Ansatz beruht grundsätzlich auf dieser Varianzmethode, verwendet aber zur Realisierung die Unabhängigkeitsanalyse. Indem sie Kovarianzen berücksichtigt, hat die Unabhängigkeitsanalyse die gewünschte Eigenschaft, dass sie die Schwankungen der Kookkurrenzhäufigkeiten der einzelnen Kontextwörter nicht isoliert betrachtet, sondern Abhängigkeiten in den Schwankungsrichtungen berücksichtigt (im genannten Beispiel also etwa zwischen *Silo* und *Getreide* sowie zwischen *Computer* und *Festplatte*). Auch ist sie nicht darauf angewiesen, dass die unterschiedlichen Bedeutungen eines Wortes in den verschiedenen Textsorten jeweils mehrheitlich vorkommen, sondern kann auch kleinere Unterschiede in den Verteilungen analysieren.

Eine andere Betrachtungsweise der Funktion der Unabhängigkeitsanalyse ist in Analogie zur in Abschnitt 1 vorgestellten Signaltrennung. Wiederum wird davon ausgegangen, dass die Bedeutungen eines mehrdeutigen Wortes in verschiedenen Textsorten typischerweise unterschiedlich häufig sind. Beispielweise möge in der einen Textsorte das Wort *Speicher* zu 90% in der Bedeutung *Computerspeicher* und zu 10% in der Bedeutung *Getreidesilo* vorkommen, während in der anderen Textsorte die entsprechenden prozentualen Anteile 70% und 30% seien. Die zu den beiden Textsorten gehörigen Kookkurrenzvektoren von *Speicher* stellen also zwei unterschiedliche Linearkombinationen derselben beiden Bedeutungen dar. Dies ist genau die Situation, die wir für die Anwendung der Unabhängigkeitsanalyse benötigen: Es liegen zwei

unterschiedliche Mischungen zweier unbekannter Bedeutungen vor. Die durch die Unabhängigkeitsanalyse bestimmten unabhängigen Komponenten dieser Mischvektoren sind maximal unterschiedliche Vektoren, die die beiden gesuchten Bedeutungen repräsentieren. Da diese Bedeutungen in der Regel nicht exakt mit Wörtern übereinstimmen werden, ist es nicht ohne weiteres möglich, sie in Worte zu fassen. In einem weiteren Schritt werden deshalb die berechneten Bedeutungsvektoren mit den auf der Basis eines Gesamtkorpus bestimmten Vektoren des gesamten Vokabulares verglichen. Die Wörter mit den höchsten Ähnlichkeiten eignen sich zur Beschreibung der Bedeutungen am besten.

4 Induktion von Wortbedeutungen mittels Unabhängigkeitsanalyse

4.1 Korpora

Folgende drei englischsprachige Korpora unterschiedlicher Textsorten wurden verwendet:

- Das *British National Corpus* (BNC), das mit der Intention, einen repräsentativen Querschnitt durch das geschriebene und gesprochene britische Englisch zu geben, aus einer Vielzahl unterschiedlicher Texte zusammengestellt wurde (ca. 100 Mill. laufende Wortformen).
- Die liberale britische Tageszeitung *The Guardian* der Jahrgänge 1992 bis 1994 (Jahrgang 1994 mit *The Observer*, insgesamt ca. 95 Mill. laufende Wortformen).
- Wissenschaftliche Abstracts der psychologischen Datenbank *PsycLIT 1988/89* und des amerikanischen *Department of Energy* (DOE). Letztere befinden sich auf der CD-ROM 1 der Data Collection Initiative der Association for Computational Linguistics (ca. 65 Mill. laufende Wortformen, etwa gleiche Anteile beider Quellen).

Da davon ausgegangen wurde, dass Funktionswörter für unsere semantischen Betrachtungen nicht von Bedeutung sind, wurden diese anhand einer etwa 200 Einträge umfassenden Stopwortliste aus den Korpora entfernt. Dadurch wurde der Speicherplatzbedarf der Korpora reduziert und die Verarbeitungsgeschwindigkeit erhöht. Zusätzlich wurden die meisten Inhaltswörter mit Hilfe eines umfangreichen Vollformenlexikons lemmatisiert (Karp et al. 1992).[3] Dadurch wird zum einen das Problem statistisch zu wenig abgesicherter Beobachtungshäufigkeiten (sparse data problem) redu-

[3] Bei unbekannten Wortformen sowie im eher seltenen Fall, dass das Lexikon zu einer mehrdeutigen Wortform mehrere mögliche Lemmata aufführte, wurde auf die Lemmatisierung verzichtet.

ziert. Zum anderen ergibt sich eine wünschenswerte Verkleinerung der zu berechnenden Kookkurrenzmatrizen.

4.2 Vergleichsdaten

Zur Evaluierung der Ergebnisse wurden die in Tabelle 1 aufgeführten zwölf mehrdeutigen Begriffe verwendet, deren Auswahl auf Yarowsky (1995) zurückgeht und für die er jeweils zwei Bedeutungen angegeben hat.

axes	grid / tools
bass	fish / music
crane	bird / machine
drug	medicine / narcotic
duty	tax / obligation
motion	legal / physical

palm	tree / hand
plant	living / factory
poach	steal / boil
sake	benefit / drink
space	volume / outer
tank	vehicle / container

Tabelle 1: Zwölf mehrdeutige Begriffe und ihre Bedeutungen nach Yarowsky (1995).

4.3 Kookkurrenzmatrizen

Wie in vielen anderen textstatistischen Untersuchungen wird zum Auszählen der Kookkurrenzhäufigkeiten zwischen Wörtern zunächst eine Fenstergröße festgelegt. Anschließend wird bestimmt, wie häufig jedes Wortpaar in einem Fenster dieser Größe gemeinsam auftritt. Die Auswahl der Fenstergröße beinhaltet einen Kompromiss zwischen den Parametern Spezifität und statistischer Schwankungsbreite. Je kleiner das Fenster gewählt wird, desto prägnanter ist in der Regel eine vorgefundene assoziative Beziehung zweier Wörter, umso stärker macht sich aber auch das Problem zufälliger statistischer Schwankungen aufgrund zu geringer beobachteter Häufigkeiten bemerkbar.

Für die vorliegende Studie wurde eine Fenstergröße von lediglich ±2 Wörtern gewählt. Diese Fenstergröße erscheint zunächst klein. Allerdings wurde die statistische Absicherung der Beobachtungen durch die Auswahl sehr umfangreicher Korpora sowie durch die vorgenommene Lemmatisierung verbessert. Zudem entspricht eine Fenstergröße von ±2 nach Eliminierung der Funktionswörter ungefähr einer Fenstergröße von ±4 bei unverändertem Text (unter der Annahme, dass im Durchschnitt jedes zweite Wort ein Funktionswort ist).

Auf der Grundlage der drei Korpora sowie eines Gesamtkorpus, das sich durch Zusammenfügen dieser Korpora ergab, wurden vier Kookkurrenzmatrizen aller in den

Korpora auftretenden ca. 700 000 unterschiedlichen Lemmata berechnet.[4] Um Worthäufigkeitseffekte zu eliminieren, wurde auf sämtliche Werte in den Matrizen der log-Likelihood-Test (Rapp 1999) angewandt. Ähnlich dem bekannteren, aber für Ereignisse mit niedriger Auftretenshäufigkeit weniger geeigneten χ^2-Test hat die Anwendung des log-Likelihood-Tests den erwünschten Effekt, dass signifikante Kookkurrenzen betont und zufällige Kookkurrenzen geschwächt werden. Zur Unterscheidung von den ursprünglichen Kookkurrenzmatrizen werden im Weiteren die so berechneten Matrizen als Assoziationsmatrizen, die zugehörigen Vektoren als Assoziationsvektoren bezeichnet.

4.4 FastICA-Algorithmus

Für die Durchführung der Unabhängigkeitsanalyse wurde der in der Programmiersprache MATLAB (MAtrix LABoratory) entwickelte FastICA-Algorithmus verwendet, der aus dem Internet unter der Adresse http://www.cis.hut.fi/projects/ica/fastica [letzter Zugriff: 15.1.2003] heruntergeladen werden kann. Eine Beschreibung der theoretischen Grundlagen der Unabhängigkeitsanalyse sowie ihrer Realisierung in Form des FastICA-Algorithmus findet sich in Hyvärinen et al. (2001).

4.5 Vektorähnlichkeiten

Die Bestimmung von Vektorähnlichkeiten ist zur Ermittlung der Bedeutungen der berechneten unabhängigen Komponenten erforderlich. Wie in früheren Arbeiten (Rapp 2002a) wird hier die City Block-Metrik verwendet, die die Ähnlichkeit zwischen zwei Vektoren X und Y als die Summe der Beträge der Differenzen korrespondierender Vektorpositionen berechnet:

$$s = \sum_{i=1}^{n} \left| X_i - Y_i \right|$$

Grundsätzlich werden alle Vektoren vor Anwendung der City Block-Metrik normalisiert, weshalb die Ergebnisse den mit rechenaufwendigeren Ähnlichkeitsmaßen (z.B. Cosinus-Koeffizient) erzielten nahe kommen dürften (vergl. Sahlgren 2002). Zur Bestimmung der zu einem vorgegebenen Assoziationsvektor bedeutungsähnlichsten anderen Wörter wird dieser Vektor mit den auf der Grundlage des Gesamtkorpus gebildeten Assoziationsvektoren aller anderen Wörter des Vokabulares verglichen und es

[4] Die Zählung der Lemmata berücksichtigt nur Zeichenfolgen, die ausschließlich aus Buchstaben bestehen.

wird eine Ähnlichkeitsrangliste erstellt. Diejenigen Wörter, für die sich die niedrigsten Werte für *s* ergeben, werden als am bedeutungsähnlichsten eingestuft.

5 Ergebnisse

Zunächst sei am Beispiel des mehrdeutigen Wortes *vessel* gezeigt, wie die Unabhängigkeitsanalyse wirkt. Schütze (1993) hat für *vessel* die drei Bedeutungen *ship*, *blood vessel* und *hollow utensil* angegeben. Abb. 5 zeigt drei Assoziationsvektoren zu *vessel*, die auf der Basis der drei Korpora berechnet wurden. Für jedes der knapp 8000 zusammen mit *vessel* auftretenden Kontextwörter ist der mit der log-Likelihood-Methode berechnete Assoziationswert aufgetragen. Aus Platzgründen sind die Wörter durchnummeriert und nur die jeweiligen Maxima beschriftet. Es ist zu erkennen, dass sowohl im British National Corpus und als auch im Guardian die Kontextwörter *blood* und *fishing* – wenn auch in unterschiedlicher Gewichtung – vorherrschen, dass diese hingegen in den wissenschaftlichen Abstracts der PsycLIT und des Department of Energy kaum eine Rolle spielen und dafür Kontextwörter wie *pressure*, *reactor* und *vacuum* im Vordergrund stehen.

Abb. 6 zeigt nun die mit Hilfe des FastICA-Algorithmus für diese drei Assoziationsvektoren berechneten unabhängigen Komponenten. Die erste unabhängige Komponente ist weitgehend identisch mit dem auf der Basis der wissenschaftlichen Abstracts berechneten Assoziationsvektor. Bei der zweiten und der dritten unabhängigen Komponente zeigt sich jedoch der bereits in Abb. 4 beobachtete Effekt: Im einen Fall wird *fishing* gestärkt und *blood* geschwächt, im anderen Fall ist es genau umgekehrt. Deutlicher als die ursprünglichen Assoziationsvektoren weisen nun also die unabhängigen Komponenten Maxima bei *pressure*, *fishing* und *blood* auf. Die Korrespondenz dieser Maxima zu den von Schütze gewählten Bedeutungsdeskriptoren ist deutlich.

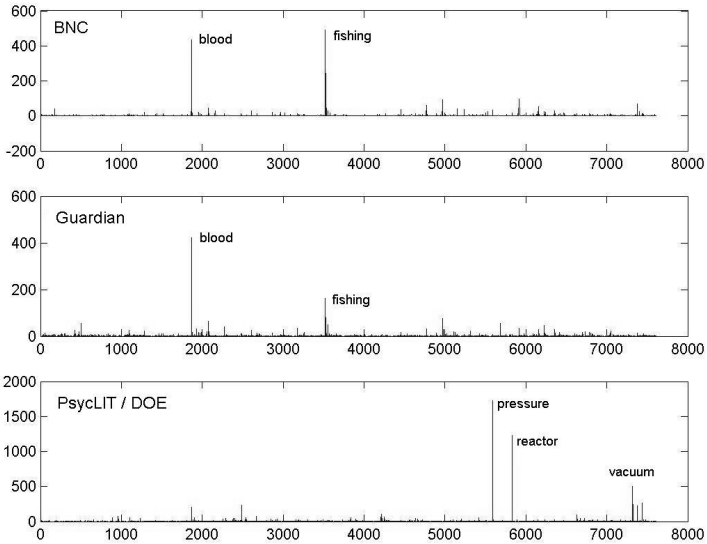

Abb. 5: Assoziationsvektoren zu *vessel* für drei Korpora.

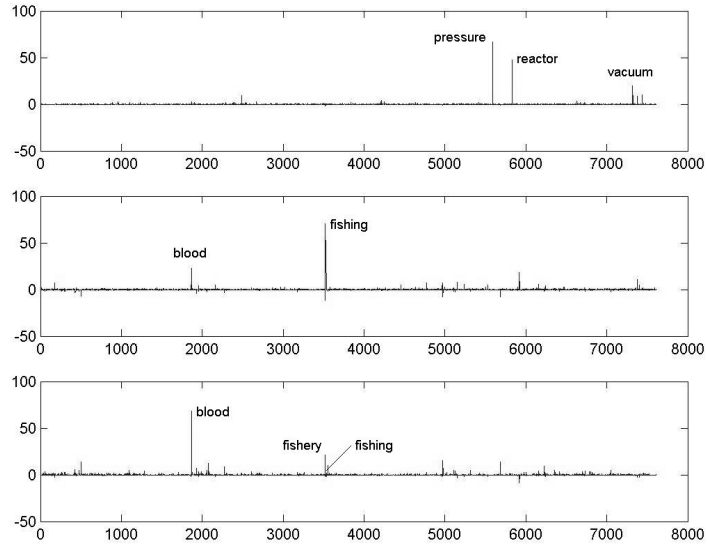

Abb. 6: Berechnete unabhängige Komponenten zu *vessel*.

Neben dieser qualitativen Betrachtung wurde auch eine quantitative Evaluierung durchgeführt. Für jedes der zwölf mehrdeutigen Wörter aus Tabelle 2 wurden auf der Basis der drei Textkorpora jeweils drei Assoziationsvektoren berechnet. In der Absicht, jeweils zwei Hauptbedeutungen zu ermitteln, wurden mit Hilfe des FastICA-Algorithmus für jedes Vektoren-Tripel zwei unabhängige Komponenten bestimmt.[5] Um die Bedeutungen der so erhaltenen unabhängigen Vektoren, die im Regelfall nicht mit den Assoziationsvektoren von Wörtern übereinstimmen, ausdrücken zu können, wurden sie mit allen Vektoren der Assoziationsmatrix des Gesamtkorpus verglichen und die jeweils drei stärksten Übereinstimmungen ausgewählt. Da die so berechneten Wörter zwar meist sehr spezifisch, häufig aber auch recht ungebräuchlich waren, wurden diejenigen Begriffe, die im Gesamtkorpus nicht mindestens eine Auftretenshäufigkeit von 100 aufwiesen, ausgesondert.[6]

Tabelle 2 gibt für jedes der zwölf mehrdeutigen Wörter nach Yarowsky (1995) diese jeweils drei zur Charakterisierung der berechneten unabhängigen Komponenten dienenden Begriffe an. Wie nicht anders zu erwarten, lassen die Ergebnisse einigen Interpretationsspielraum. Nach unserer Beurteilung stimmen die Ergebnisse für *axes, bass, drug, motion, plant* und *space* recht deutlich mit den von Yarowsky vorgegebenen Kategorien überein. Mit Abstrichen gilt dies auch für *palm* und für *tank*, insbesondere wenn man berücksichtigt, dass die Wortformen *cupping* und *cupped* fast ausschließlich zusammen mit *hand* auftreten (cupped hand = hohle Hand),[7] und dass die Abkürzung *SMES* in den Abstracts des Department of Energy für *superconducting magnetic energy storage* steht. Im Falle von *duty* und *sake* wurden andere als die von Yarowsky angegebenen Unterscheidungen vorgenommen, die aber dennoch plausibel erscheinen. Bei *duty* sind dies die Bedeutungen im Sinne von *obligation* und von *heavy duty*; bei *sake* die Verwendungen im Sinne von *sake of God* und von *sake of clarity* (in Texten). Zumindest bei *sake* ist recht klar, warum die von Yarowsky vorgegebene Unterscheidung nicht gefunden werden konnte, denn die Bedeutung im Sinne des

[5] Bei der Unabhängigkeitsanalyse kann die Zahl der zu berechnenden unabhängigen Komponenten ohne weiteres auch kleiner sein als die Zahl der zur Verfügung stehenden Mischvektoren. Eine höhere Anzahl ist jedoch nicht möglich. Da der FastICA-Algorithmus iterativ arbeitet und von einer zufälligen Initialisierung ausgeht, kann es gelegentlich je nach Initialisierung zu unterschiedlichen Ergebnissen kommen. Die hier vorgestellten Ergebnisse beziehen sich immer auf den ersten Lauf.

[6] Anstatt eine Mindesthäufigkeit vorzugeben, könnte die Kandidatenauswahl auch aufgrund der berechneten Assoziationsstärke zum mehrdeutigen Wort erfolgen. Beispielsweise wäre es möglich, nur die stärksten 100 Assoziationen zu berücksichtigen.

[7] Die Wortformen *cupped* und *cupping* wurden deshalb nicht lemmatisiert, weil ohne Kontextanalyse ihre Grundformen nicht eindeutig sind. Bei beiden ist zwar die Rückführung auf das Verb *to cup* möglich, bei *cupped* gibt es aber zudem den Gebrauch als Adjektiv und bei *cupping* den als Substantiv (Grundformen jeweils unverändert).

japanischen Reisweines dürfte in den verwendeten Korpora allenfalls äußerst selten vorkommen. In den restlichen Fällen, nämlich bei *crane* und *poach*, weist zwar jeweils mindestens eine Spalte eine gute Übereinstimmung mit einer der vorgegebenen Bedeutungen auf. Es war uns jedoch nicht möglich, der jeweils anderen Spalte eine klare Bedeutung zuzuordnen, die sich davon deutlich abheben würde. Für unsere allerdings sehr kleine Stichprobe von zwölf Wörtern ergab sich somit zu etwa 67% das erwartete Ergebnis, zu knapp 17% zwar nicht genau das erwartete, aber dennoch ein akzeptabel erscheinendes Ergebnis, und zu ebenfalls 17% ein fehlerhaftes Ergebnis.

axes grid / tools		bass fish / music		crane bird / machine	
mach	axis	catfish	guitar	manipula	winch
ete	crystallograph	sturgeon	saxophon	tor	truck
axe	ic	minnow	e	monorail	bulldoz
crow	orthogonal		solo	equipme	er
bar				nt	

drug medicine / narcotic		duty tax / obligation		motion legal / physical	
tranquiliz	medicati	obligation	motorcycl	velocity	amendme
er	on	Customs	es	nonline	nt
heroin	polydrug	responsibili	drinker	ar	vote
cocaine	inhalant	ty	snowfall	trajecto	MPs
				ry	

palm tree / hand		plant living / factory		poach steal / boil	
shrub	cupping	hydroelectric	shrub	boil	tab
trunks	cupped	Bonneville	tree	omelette	LSP
felling	outstretch	PowerGen	flower	yolk	app

sake benefit / drink		space volume / outer		tank vehicle / container	
almighty	brevity	Nasa	correspon	retrievabl	artiller
bless	cheapness	astronaut	d	e	y
crucify	meaningful	laboured	dimensio	SMES	howitz
			nal	pressurize	er
			surface	r	infantr
					y

Tabelle 2: Ergebnisse für die mehrdeutigen Wörter nach Yarowsky (1995).

6 Zusammenfassung und Ausblick

Am Beispiel der Induktion von Wortbedeutungen wurde gezeigt, dass sich die Unabhängigkeitsanalyse dafür eignet, das bisher weitgehend ungelöste Problem der Mehrdeutigkeit natürlicher Sprache anzugehen. Die erhaltenen Ergebnisse deuten darauf hin, dass die Unabhängigkeitsanalyse in der Lage ist, auf Grund von Unterschieden in der Verwendung mehrdeutiger Begriffe in unterschiedlichen Textsorten die Wortbedeutungen zu erschließen,[8] wobei vorausgesetzt wird, dass die einzelnen Bedeutungen in unterschiedlichen Textsorten verschieden häufig auftreten.[9] Diese Voraussetzung dürfte allerdings weniger grundsätzlicher Natur sein, sondern in erster Linie eine Frage der Korpusgröße und -anzahl. Bislang wurden aus Gründen der Verfügbarkeit allerdings lediglich drei Textsorten verwendet. Naheliegend und psychologisch plausibler wäre es, eine viel größere Anzahl von Textsorten zu unterscheiden, oder gar jedes einzelne Dokument als eigene Textsorte zu betrachten. Letzteres wäre eine neue Umsetzung der von Yarowsky (1995) prägnant als „one sense per discourse“ formulierten Beobachtung, dass mehrdeutige Wörter im selben Dokument meist nur in einer Bedeutung verwendet werden.

Dem entgegen steht jedoch das Problem, dass jedes Dokument nur einen sehr kleinen Teil des in einer Sprache vorhandenen Wortschatzes enthält, und dass die Überschneidungen zwischen verschiedenen Dokumenten deshalb gering sind. Dies gilt, wenngleich in geringerem Maße, auch für thematisch verwandte Dokumente, da selbst gleichartige Sachverhalte je nach Autor häufig mit einem unterschiedlichen Vokabular ausgedrückt werden.

Eine mangelnde Überlappung der Vokabulare führt jedoch dazu, dass die Unabhängigkeitsanalyse keine brauchbaren Ergebnisse liefern kann. Eine in dieser Arbeit bereits angewandte Vorgehensweise, dieses Problem zu verringern, bestand darin, durch die Lemmatisierung der Korpora die Vielzahl vorkommender Wortformen auf

---

[8] Im Falle von Mehrdeutigkeiten, die sich besser über die Syntax auflösen lassen (etwa die Bedeutungen des Wortes *einen* als Verb und als Artikel), sollten allerdings Kookkurrenzinformationen zugrunde gelegt werden, die die Wortreihenfolge berücksichtigen, was die Einbeziehung wichtiger syntaktischer Informationen bedeutet (vergl. Rapp 2002b).

[9] Ein alternativer Ansatz bestünde darin, nicht die unabhängigen Komponenten zu verschiedenen Kookkurrenzvektoren desselben Wortes zu berechnen, sondern – auf der Basis nur eines Korpus – diejenigen zwischen dem mehrdeutigen Wort und seinen stärksten Assoziationen. Wie Vorversuche ergaben, treten hierbei allerdings zwei Schwierigkeiten auf: Zum einen sind die Assoziationen meist ebenfalls mehrdeutig, so dass nicht klar ist, ob eine berechnete unabhängige Komponente eine Bedeutung des Ausgangswortes oder eine sonstige Bedeutung einer Assoziation repräsentiert. Zum anderen kommt es bei mehrdeutigen Wörtern häufig vor, dass die Hauptbedeutung so stark überwiegt, dass unter den stärksten Assoziationen eine weniger häufige Bedeutung gar nicht vertreten ist.

eine etwas geringere Anzahl von Lemmata zu reduzieren. Nötig ist jedoch eine viel stärkere Reduktion.

Diese kann unseres Erachtens mit Hilfe der *Singular Value Decomposition* (SVD) erreicht werden (Press et al. 1992; Landauer & Dumais 1997; Manning & Schütze 1999). Die SVD ermöglicht es, die Dimensionalität (Anzahl der Spalten) der betrachteten Assoziationsmatrizen und -vektoren in optimaler Weise zu reduzieren, d.h. so, dass sich an den euklidischen Distanzen zwischen den einzelnen Zeilenvektoren möglichst wenig ändert.

Eine drastische Reduktion der Anzahl der zu berücksichtigenden Spalten auf die z.B. 200 wichtigsten Dimensionen dürfte die Anwendbarkeit der Unabhängigkeitsanalyse stark verbessern. Inwieweit hierbei allerdings die Gefahr besteht, dass für die Unterscheidung ähnlicher Bedeutungen nötige Informationen verloren gehen, können erst künftige Experimente zeigen.

Zum Schluss sei darauf hingewiesen, dass sich die hier vorgestellte Vorgehensweise auch auf viele andere Problemstellungen übertragen lässt, bei denen die Überlagerung von Vektoren eine Rolle spielt (Rapp 2002b), beispielsweise solchen aus den Bereichen Information Retrieval, Wort- und Dokumentklassifikation sowie Memory-Based Learning (Daelemans et al. 2002).

Danksagungen

Ich danke Manfred Wettler und Gisela Zunker-Rapp für die inhaltliche und der Deutschen Forschungsgemeinschaft für die finanzielle Unterstützung dieser Arbeit. Mein Dank gilt auch Erkki Oja für seinen wegweisenden Vortrag bei der 26. Jahrestagung der Gesellschaft für Klassifikation 2002 in Mannheim sowie Jarmo Hurri, Hugo Gävert, Jaakko Särelä und Aapo Hyvärinen für die Bereitstellung des FastICA-Algorithmus.

Literatur

Arns, Ursula (1994). Sprachstatistische Analysen lexikalischer Mehrdeutigkeiten. Diplomarbeit an der Universität-GH Paderborn, Fachbereich Psychologie.

Daelemans, Walter; Zavrel, Jakub; van der Sloot, Ko; van den Bosch, Antal (2002): *TiMBL: Tilburg Memory Based Learner, Version 4.2, Reference Guide.* ILK Technical Report 02-01. http://ilk.kub.nl

Hyvärinen, Aapo; Karhunen, Juha; Oja, Erkki (2001): Independent Component Analysis. New York: Wiley.

Karp, Daniel; Schabes, Yves; Zaidel, Martin; Egedi, Dania (1992): A freely available wide coverage morphological analyzer for English. In: Proceedings of the 14th International Conference on Computational Linguistics. Nantes, Frankreich, 950-955.

Landauer, T. K.; Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211-240.

Lin, Dekang; Pantel, Patrick (2002). Concept discovery from text. In: Proceedings of the International Conference on Computational Linguistics (COLING), Taipei, Vol. 1, 577–583.

Manning, Christopher D.; Schütze, Hinrich (1999): Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press.

Press, William H.; Teukolsky, Saul A.; Vetterling, William T.; Flannery, Brian P. (1992): Numerical Recipes in C. The Art of Scientific Computing. 2nd edition. Cambridge University Press.

Rapp, Reinhard (1999): Automatic identification of word translations from unrelated English and German corpora. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, Maryland. 519–526.

– (2002a): The computation of word associations: comparing syntagmatic and paradigmatic approaches. Proceedings of the 19th International Conference on Computational Linguistics, Taipei, ROC, Vol. 2, 821-827.

– (2002b); Unsupervised learning of second order dependencies from corpora. Tagungsband der 6. Konferenz zur Verarbeitung Natürlicher Sprache, DFKI, Saarbrücken.

Roberts, Stephen; Everson, Richard (Hrg.) (2001). Independent Component Analysis. Principles and Practice. Cambridge University Press.

Sahlgren, Magnus (2002): Vector-based semantic analysis: representing word meanings based on random labels. In: A. Lenci, S. Montemagni, V. Pirrelli (Hrg.): Acquiring and Representing Semantic Knowledge. Dordrecht: Kluwer.

Schütze, Hinrich (1993): Word space. In: S.J. Hanson, J.D. Cowan, C.L. Giles (Hrg.): Advances in Neural Information Processing Systems 5, 895-902. San Mateo CA: Morgan Kaufman Publishers.

Yarowsky, David (1995): Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, MIT Cambridge, 189–196.