Anwendungen des semantischen Wissens über Konzepte im *Information Retrieval*

Dr. Iryna Gurevych
Natural Language Processing Gruppe
EML Research gGmbH
Schloss-Wolfsbrunnenweg 33
69118 Heidelberg, Germany

Abstract (Deutsch)

Die im Folgenden beschriebenen experimentellen Arbeiten befassen sich mit der Integration des semantischen Wissens über Konzepte in das *Information Retrieval* (IR). Verschiedene Arten des Wissens werden aus dem lexikalisch-semantischen Wortnetz GermaNet extrahiert. Einerseits wurde geprüft, ob die Performanz eines IR-Systems durch eine Erweiterung der Anfrage mit Wörtern aus automatisch generierten Definitionen der Konzepte verbessert wird. Andererseits wurde ein neuartiges *Information Retrieval* Modell getestet, dem lexikalisch-semantische Verwandtschaft der Wörter zugrunde liegt. Die Ergebnisse aller Experimente wurden auf einem IR-Datensatz in der Domäne „Elektronische Berufsberatung“ mit einem herkömmlichen IR-System verglichen und automatisch ausgewertet. Dabei ergab sich: 1) eine Erweiterung der Anfragen durch Hyponyme ist besonders nützlich; 2) das semantische IR-Modell schneidet auf der gleichen Ebene ab, wie das tf*idf IR-Modell.

Abstract (Englisch)

The experimental work described in this paper investigates the integration of the semantic information about concepts in Information Retrieval (IR). Different types of knowledge are extracted from the lexical-semantic database GermaNet. We studied the performance of an IR system, where the query is expanded with words from automatically generated definitions of concepts, and tested a novel information retrieval model, which is based on lexical-semantic relatedness of words. The results of our experiments were evaluated and compared with a state-of-the-art system, using an IR test collection in the domain “Electronic Career Consultancy”. We found out the following: 1) using hyponyms for the query expansion is definitely beneficial; 2) the semantic IR model performs at the same level as the tf*idf based IR system.

1. Einführung

Semantisches Wissen wird oft für die Verbesserung von *Information Retrieval* Systemen benötigt. Die Forschungshypothese ist, dass dieser Typ des Wissens die Qualität gegenwärtiger *Information Retrieval* Systeme erhöhen wird, denen statistische Methoden und Zeichenkettenvergleiche zugrunde liegen. Bisherige Forschungsarbeiten konnten das nicht überzeugend und konsistent beweisen (vgl. z.B. Evens, 2002).
Für die Einbindung semantischen Wissens in das *Information Retrieval* existieren verschiedene Möglichkeiten:

- Anfrageerweiterung. Hierbei werden die in der Anfrage enthaltenen Terme mit ihren semantisch verwandten Begriffen erweitert. Als Ergebnis können Dokumente gefunden werden, die den ursprünglichen Anfrageterm nicht enthalten, jedoch mit ihm semantisch verwandte Begriffe. Wichtig ist hierbei die Metho-

de, mit deren Hilfe semantisch verwandte Begriffe für einen Anfrageterm bestimmt werden (Voorhees, 1994; Mandala et al., 1998).

- Konzepte statt Wörter indexieren. Hierbei wird der Index nicht auf der Wortebene erzeugt, sondern auf der Konzeptebene. Wörter werden in der Vorverarbeitung auf Begriffe abgebildet, so dass Anfragen dann gegen einen Konzept-Index abgeglichen werden (Gonzalo et al., 1998; Voorhees, 1999).
- Semantische Verwandtschaft als Modell des *Information Retrievals.* Bei diesem Verfahren werden klassische *Information Retrieval* Modelle durch ein alternatives Modell ersetzt. Das alternative Modell berechnet die Relevanz eines Dokumentes in bezug auf eine Anfrage aufgrund der semantischen Verwandtschaft der darin enthaltenen Wörter (Smeaton, 1999).

In diesem Beitrag beschreiben wir Ergebnisse unserer Pilot-Studie, die sich semantischen Wissens über Begriffe in einem *Information Retrieval* System bedient. Das System greift auf das semantische Wissen in GermaNet zu, dem lexikalisch-semantischen Wortnetz für die deutsche Sprache (Kunze, 2004). In diesem Wortnetz sind Nomen, Verben und Adjektive als *is-a* Hierarchien repräsentiert. Zusätzliche Informationen auf lexikalischer (Wort-) Ebene, z.B. Synonymie, Antonymie, Derivation, und semantischer (Wortbedeutungs-) Ebene, z.B. Meronymie, Assoziation können sprachverarbeitenden Anwendungen zur Verfügung gestellt werden.

Zwei Serien von Experimenten wurden durchgeführt. In der Serie haben wir die Anfrage des Benutzers mit verwandten Begriffen erweitert. Für die Anfrageerweiterung wurden verschiedene Arten der aus GermaNet erzeugten künstlichen Definitionen der Wortbedeutungen (Konzepte) herangezogen. Automatisch generierte Definitionen wurden zum ersten Mal für die Berechnung semantischer Verwandtschaft von zwei Wörtern eingeführt (Gurevych, 2005b). Wichtige Parameter für die Erzeugung der Definitionen sind Typen der semantischen Beziehungen in GermaNet, z.B. Hyperonymie, Synonymie, Assoziation und der Verwandtschaftsgrad (Distanz in Kanten vom Konzept zu verwandten Konzepten). Für die Berechnung semantischer Verwandtschaft zwischen zwei Konzepten hat sich die Hyperonymie-Beziehung bis zum Verwandtschaftsgrad 3 als besonders hilfreich erwiesen. In der vorliegenden Arbeit galt es herauszufinden, welche Parameter für die Erzeugung künstlicher Definitionen der Konzepte im *Information Retrieval* Kontext besonders hilfreich sind.

Für die zweite Serie von Experimenten wurden verschiedene Maße der semantischen Verwandtschaft entwickelt und evaluiert. Semantische Verwandtschaft wurde als eine beliebige Art der semantischen oder assoziativen Beziehung zwischen zwei Begriffen definiert (Gurevych & Niederlich, 2005b). Ein Maß semantischer Verwandtschaft (Lin, 1998) wurde im semantischen *Information Retrieval* für die Bestimmung der relevanten Dokumente benutzt. Aus diesem Verfahren resultierende Ergebnisse wurden mit denen eines konventionellen IR Modells verglichen und anhand eines *Gold Standards* evaluiert.

Die in dieser Arbeit beschriebenen Methoden wurden mit dem Wortnetz GermaNet erprobt, können jedoch auf beliebige konzeptuelle Netzwerke und Taxonomien über-

tragen werden. Insbesondere sind für fachspezifische Gegenstandsbereiche, wie z.B. Bibliothekwissenschaften oder Biomedizin, solche Ressourcen oft bereits vorhanden und auf spezielle Domänen zugeschnitten. In diesem Fall kann auch eine fachspezifische Ressource, die domänenspezifisches Wissen im großen Umfang und Detail repräsentiert, statt einer allgemeinsprachlichen Ressource eingesetzt werden.
Das in dieser Arbeit beschriebene *Information Retrieval* System wurde in der Domäne „Elektronische Berufsberatung" erprobt und evaluiert. In dieser Anwendung werden Aufsätze mit Beschreibungen beruflicher Interessen von Personen (Interessenprofile) als Anfragen aufgefasst und gegen eine Dokumentenkollektion mit natürlichsprachlichen Beschreibungen der Ausbildungsberufe automatisch abgeglichen.[1] Ein Interessenprofil gegeben, liefert das System eine nach Relevanz geordnete Liste der Berufe zurück. Die Performanz des Systems wird mit Hilfe der TREC Evaluierungsmetriken gemessen.[2] Der *Gold Standard* für die Evaluierung wurde durch ein wissensbasiertes System simuliert (Gurevych, 2005a), dem die Datenbank „Interesse:Beruf" der Bundesagentur für Arbeit zugrunde liegt.[3] Das Evaluierungsmaß berechnet gemittelte Präzision für alle relevanten Dokumente (gemittelt über Anfragen).
Das vorliegende Papier wird die folgende Struktur haben: In Kapitel 2 wird die Anwendungsdomäne „Elektronische Berufsberatung" näher beschrieben. Insbesondere werden die Aufgabe und die in Experimenten eingesetzten Testdaten charakterisiert. Kapitel 3 stellt dann die experimentellen Arbeiten detailliert vor. Künstliche Definitionen der Konzepte und Maße semantischer Verwandtschaft werden erläutert. Anschließend beschreiben wir die Einbindung dieser Wissensarten in das *Information Retrieval*. Experimentelle Ergebnisse werden vorgestellt und diskutiert. Im letzten Kapitel 4 fassen wir den Stand der Arbeiten knapp zusammen und erarbeiten ausgehend von den Ergebnissen Richtlinien für zukünftige Forschungsarbeiten.

2. Elektronische Berufsberatung

In der vorliegenden Arbeit wurde elektronische Berufsberatung als Anwendungsdomäne für das semantische *Information Retrieval* gewählt. „Berufsberatung ... wird meistens von Schülern in Anspruch genommen um zu erfahren, welcher Beruf zu ihnen passt, welche Anforderungen und Kenntnisse gefordert werden. Aber auch Arbeitslose, die aus gesundheitlichen Gründen arbeitslos geworden sind oder im erlernten Beruf keine Perspektive mehr sehen, informieren sich über Umschulungsmöglichkeiten".[4] (Wikipedia, 2005)
Personen, die auf der Suche nach einem zu ihnen passenden Beruf sind, können in der Regel von einem speziell ausgebildeten Experten beraten werden. Ein Berufsberater besitzt umfangreiches Domänenwissen über die Berufswelt. Diese Art der Berufsberatung ist jedoch teuer (Personalkosten), nicht immer zugänglich (Öffnungs-

[1] Bundesagentur für Arbeit. BERUFENet. http://berufenet.arbeitsamt.de/, Nürnberg, Germany.
[2] TREC. http://trec.nist.gov/overview.html
[3] Bundesagentur für Arbeit. Interesse:Beruf. http://www.interesse-beruf.de/, Nürnberg, Germany.
[4] http://de.wikipedia.org/wiki/Berufsberatung

und Wartezeiten), und schlecht reproduzierbar (unterschiedliche Berater geben verschiedene Empfehlungen aus).
Auf der anderen Seite stehen den Ratsuchenden automatisierte Berufsberatungsmöglichkeiten wie z.B. das bereits erwähnte Programm „Interesse:Beruf“ zur Verfügung. Dabei wird der Benutzer gebeten, die auf ihn oder sie zutreffenden Schlagwörter aus drei Kategorien (Was? Wo? Womit?) auszuwählen. Das System greift dann auf eine Datenbank zu, in dem zu jedem Beruf relevante Schlagwörter von Fachexperten vergeben wurden, und berechnet zu jedem Beruf die Anzahl der Treffer. Eine nach Anzahl der Treffer geordnete Liste der Berufe wird dem Benutzer zurückgeliefert. Diese Art der Berufsberatung ist für den Benutzer jederzeit über ein Web-Interface zugänglich und die Ergebnisse sind, zumindest für die jeweilige Version der Datenbank, reproduzierbar. Nichtsdestotrotz ist die Freiheit des Benutzers in der Beschreibung seiner persönlichen Interessen durch eine fest vorgegebene Menge der Schlagwörter stark reduziert. Ein weiterer Nachteil dieser Methode ist der erhebliche manuelle Aufwand für die Pflege der Datenbank mit Zuordnungen einzelner Berufe und Schlagwörter. Oft werden neue Berufe geschaffen, und noch öfter verändert sich die Beschreibung eines bestehenden Berufes, so dass die Schlagwörter geändert werden müssen.
Im vorliegenden Papier schlagen wir einen alternativen, sprachbasierten Zugang zu der Berufsberatung vor. Der Zugang wird als sprachbasiertes Beratungssystem implementiert. Die Aufgabe der Berufsberatung wird als eine *Information Retrieval* Aufgabe definiert. Gegeben eine Anfrage (Interessenprofil), sollen Dokumente (Berufsbeschreibungen) nach ihrer Relevanz für die Anfrage geordnet werden. Ein derartiges System würde dem Benutzer ermöglichen, die Berufsberatung jederzeit in Anspruch zu nehmen und dabei frei in der sprachlichen Formulierung beruflicher Interessen und Vorstellungen zu bleiben. Für die Pflege der Datenbank fällt kein zusätzlicher Aufwand an. Benötigt werden lediglich ausführliche Beschreibungen der einzelnen Berufe, die in der Regel bereits vorliegen.
Für die Evaluierung eines *Information Retrieval* Systems wird ein Testdatensatz benötigt, der folgende Komponenten enthält: eine Kollektion von Dokumenten, eine Kollektion von Themen,[5] und eine Menge von Relevanzurteilen für jedes Thema in bezug auf alle Dokumente, die *Gold Standard* genannt wird.
Die Domäne der elektronischen Berufsberatung eignet sich wegen ihrer speziellen Eigenschaften besonders gut für unsere Experimente. Auf der einen Seite liegt eine Dokumentenkollektion mit den Beschreibungen von etwa 1.800 Ausbildungsberufen, z.B. Altenpfleger/in, Elektroniker/in, und weiteren 4.000 Berufen vor, z.B. Informatiker/in, Maschinenbauingenieur/in, vor. Diese Berufsbeschreibungen werden in der Datenbank BERUFENet verwaltet, die in der Einführung bereits erwähnt wurde. In-

[5] Der Begriff „Thema“ soll im Kontext des *Information Retrievals* vom Begriff „Anfrage“ unterschieden werden. Mit dem Thema wird eine natürlichsprachliche Beschreibung des Informationsbedürfnisses eines Benutzers gemeint. Hingegen wird unter „Anfrage“ eine Menge von Suchtermen verstanden, auf die das ursprüngliche Thema als Folge einer automatischen Vorverarbeitung abgebildet wurde.

formationen über einzelne Berufe werden im XML-Format repräsentiert. Das Datenmodell sieht ca. 60 verschiedene Datenfelder vor, wie z.B. Inhalte der Berufsausbildung, Aufgaben und Tätigkeiten im jeweiligen Beruf, Angaben zu erwarteten Kompetenzen eines Bewerbers. Auf der anderen Seite liegt die Datenbank „Interesse:Beruf“ vor. In (Gurevych, 2005a) wurde ein Verfahren vorgestellt, das einen *Gold Standard* für die Evaluierung der Ergebnisse des *Information Retrievals* aus dieser Datenbank automatisch generiert. Dafür sollen lediglich die Themen, d.h. Interessenprofile, mit zutreffenden Schlagwörtern aus der Datenbank annotiert werden. Es ist nicht mehr nötig, einzelne Dokumente im Hinblick auf ihre Relevanz zur Anfrage manuell zu beurteilen. Das wäre nicht nur aufwendig, sondern würde spezielle Fachkompetenz auf dem Gebiet der Berufsberatung erfordern.
Da „Interesse:Beruf“ für 578 Ausbildungsberufe in Deutschland entwickelt wurde, setzt sich unsere Testkollektion aus den Beschreibungen dieser Berufe zusammen. Um einige Beispiele für die Eingaben in das System zu sammeln, wurden Interessenprofile in einem Experiment mit 30 Versuchspersonen erhoben. Die Teilnehmer am Experiment haben kurze Texte über ihre beruflichen Interessen und Erwartungen geschrieben (Beispiel 1), die eine Eingabe in das IR-System darstellen:

(1) „Ich würde gerne mit Tieren arbeiten, sie behandeln, für sie sorgen, aber ich kann kein Blut sehen und ich habe zu viel Mitleid mit kranken Tieren. Andererseits arbeite ich besonders gerne am Computer, kann programmieren in C, Python und VB und könnte mir daher auch in der Software-Entwicklung einen passenden Beruf vorstellen. Ich kann mir nur schlecht vorstellen in einem Kindergarten, als Sozialberater oder als Lehrer zu arbeiten, da ich mich nicht besonders gut durchsetzen kann.“

3. Semantisches Wissen im *Information Retrieval*

3.1 Baseline Information Retrieval System

Die Ergebnisse des semantischen *Information Retrievals* werden im Folgenden mit einem herkömmlichen Baseline-System verglichen. Das implementierte Baseline-System basiert auf dem erweiterten booleschen Modell (Salton et al., 1983). Im ersten Schritt werden die Dokumente in der Testkollektion unter der Verwendung einer allgemeinen deutschen Stoppwortliste und des *Stemming* indexiert. Im zweiten Schritt werden Anfragen gegen den Index abgeglichen und Dokumente in relevante und irrelevante eingeteilt. Die relevanten Dokumente werden anschließend nach der Gleichung 1 geordnet.

$$\sum_{t\_in\_d} tf(t\_in\_d) \times idf(t)$$

Gleichung 1

tf(t_in_d) ist der Termfrequenzfaktor des Termes (*t)* im Dokument *(d)*, und *idf(t)* ist die inverse Dokumentenfrequenz des Termes.

3.2 Experimente mit der Anfrageerweiterung

Eine Erweiterung der Anfrage im *Information Retrieval* wird dadurch motiviert, dass Eingaben der Benutzer auf einer Seite und Dokumente in der Kollektion auf der anderen Seite oft ein unterschiedliches Vokabular, d.h. Wortschatz, benutzen. In der Berufsberatung werden Interessenbeschreibungen eher informell formuliert. Berufsbeschreibungen weisen dagegen eine formale Ausdrucksweise auf. Für den gleichen Begriff werden verschiedene Wörter benutzt, z.B. kann der Benutzer das Wort „Brötchen“ verwenden, wobei „Backwaren“ in der Berufsbeschreibung auftritt.

Dokumente werden als Erstes unter Verwendung der Java-basierten IR-Bibliothek Lucene indexiert.[6] Interessenprofile werden vorverarbeitet (Tokenisierung, Wortartenerkennung), und entweder als eine Menge von Nomen (N) oder als eine Menge von Inhaltswörtern (Nomen, Verben, Adjektive, Adverbien - NVAA) repräsentiert. Daraus resultierende Anfragen an das IR-System werden unter Zugriff auf die Java-basierte GermaNet API mit den Begriffen aus automatisch erzeugten Definitionen expandiert. Dazu wurden verschiedene Typen der semantischen Relationen mit unterschiedlichem Verwandtschaftsgrad verwendet. Auf diesen Repräsentationen wurde das in Lucene implementierte IR-Verfahren (s. Abschnitt 3.1) angewandt und die Ergebnisse mit Hilfe der TREC Evaluierungssoftware ausgewertet (Tabelle 1).

Typ sem. Beziehung	Verwandtschaftsgrad	N	Differenz	NVAA	Differenz
Antonymie	1	0,3153	-0,0034	0,3204	-0,0161
	3	0,3153	-0,0034	0,3204	-0,0161
	alle	0,3153	-0,0034	0,3204	-0,0161
Holonymie	1	0,3238	0,0051	0,3411	0,0046
	3	0,3194	0,0007	0,3378	0,0013
	alle	0,3194	0,0007	0,3378	0,0013
Hyperonymie	1	0,3250	0,0063	0,3280	-0,0085
	3	0,3070	-0,0117	0,3378	0,0013
	alle	0,3163	-0,0024	0,3224	-0,0141
Hyponymie	1	**0,3480**	**0,0293**	**0,3530**	**0,0165**
	3	**0,3459**	**0,0272**	**0,3509**	**0,0144**
	alle	**0,3492**	**0,0305**	**0,3544**	**0,0179**
Meronymie	1	0,3188	0,0001	0,3386	0,0021
	3	0,3173	-0,0014	0,3368	0,0003
	alle	0,3173	-0,0014	0,3368	0,0003
Synonymie	1	0,3100	-0,0087	0,3235	-0,0130
	3	0,3037	-0,0150	0,3175	-0,0190
	alle	0,3100	-0,0087	0,3208	-0,0157
Alle	alle	**0,3471**	**0,0284**	**0,3586**	**0,0221**
Baseline		0,3187		0,3365	

Tabelle 1

[6] http://lucene.apache.org/java/docs/

Bei den beiden Arten der Anfragerepräsentation (N oder NVAA) kommt es zu ähnlichen Ergebnissen. Die Ergebnisse für die letztere Systemkonfiguration sind allgemein etwas besser, als für die erste. Bei der Bestimmung der optimalen Parameter zur Erzeugung der künstlichen Definitionen wird deutlich, dass lediglich Hyponyme mit steigendem Verwandtschaftsgrad eine Verbesserung des *Information Retrieval* hervorrufen. Z.B. wird das Konzept „Computer“ durch folgende Begriffe erweitert: „Laptop Notebook Mainframe Minicomputer PC Workstation Macintosh Apple Client Gateway Server“. Der in Lucene implementierte *Stemming*-Mechanismus verursacht einige Fehler. So wird „Blut“ auf den Stamm „blu“ abgebildet, und dieser wird dann nicht nur mit dem ursprünglichen „Blut“, sondern auch mit der „Bluse“ assoziiert. Als Folge wird die Anfrage teilweise mit irrelevanten Begriffen erweitert, z.B. „Top Hemdbluse Seidenbluse Baumwollbluse Leinenbluse Röschenbluse Blutkonserve Konserve“.
Ein weiterer Nachteil der bestehenden Systemarchitektur ist, dass keine Wortlesartendisambiguierung vorgenommen wird. Das bedeutet, dass die Anfrage nicht mit der Definition der aktuellen Bedeutung eines Wortes erweitert wird, sondern mit den Definitionen aller seinen Bedeutungen. Relevante Forschungsarbeiten haben bisher gezeigt, dass eine Disambiguierung der Wortlesarten keine bemerkbaren Verbesserungen des *Information Retrievals* geleistet hatte (vgl. Sanderson (1994). Der Grund ist insbesondere, dass die Wortlesartendisambiguierung perfekt sein muss, um positive Auswirkungen auf IR zu haben. Die aktuelle Situation auf diesem Gebiet ist jedoch, dass die Genauigkeit der Algorithmen etwa 60-70% beträgt. Diese können somit nicht mit Gewinn in das sprachbasierte Beratungssystem integriert werden.

3.2 Semantische Verwandtschaft der Wörter

In gegenwärtigen *Information Retrieval* Systemen wird die Relevanz eines Dokuments in Bezug auf eine Anfrage auf der Grundlage des booleschen, probabilistischen oder des Vektorraummodells bestimmt. Anfragen und Dokumente werden als Mengen von Index-Termen repräsentiert. Zwischen den einzelnen Wörtern existierende lexikalische und semantische Relationen werden nicht berücksichtigt. Dadurch werden relevante Dokumente, die ein anderes Vokabular benutzen als die Anfrage, nicht gefunden. Eine Alternative zu zeichenkettenbasierten Verfahren des *Information Retrieval*s ist eine Approximierung der Relevanz eines Dokuments in bezug auf eine Anfrage mit Hilfe von Maßen lexikalisch-semantischer Ähnlichkeit. Das bedeutet, dass die relevantesten Dokumente der Anfrage semantisch am ähnlichsten sind. Die Zwischenrepräsentation für das *Information Retrieval* wird verbessert, indem Wörter der natürlichen Sprache auf lexikalische Konzepte in GermaNet abgebildet werden. Die Bestimmung semantischer Ähnlichkeit bezieht das in GermaNet modellierte lexikalische Wissen, Domänen- und Weltwissen sowie die Ergebnisse einer umfassenden Korpusanalyse mit ein.
Existierende Ansätze zur Berechnung semantischer Verwandtschaft können in drei verschiedene Klassen eingeteilt werden: *wörterbuchbasierte* Maße (Lesk, 1986; Gu-

revych, 2005b), *distanzbasierte* Maße (Hirst & St-Onge 1998; Leacock & Chodorow, 1998), und *informationsgehaltsbasierte* Maße (Resnik, 1995; Lin, 1998).
Wörterbuchbasierte Maße beruhen auf der Annahme, dass in Wörterbuchdefinitionen semantisch verwandter Wörter viele Wortüberlappungen zu finden sind. Die in unseren Experimenten benutzte Ressource, GermaNet, enthält nur wenige Definitionen. Dagegen findet man in einem semantischen Netzwerk semantisch verwandte Konzepte. In Gurevych (2005b) wurde ein neues Verfahren vorgeschlagen, das diese Eigenschaft des Netzwerks ausnutzt und künstliche Definitionen eines Konzeptes nach vorgegebenen Parametern generieren kann. Die Wortüberlappung für die Bestimmung semantischer Verwandtschaft wird dann nicht auf echten, sondern auf automatisch generierten Definitionen berechnet. Dies ermöglicht die Bestimmung semantischer Verwandtschaft zwischen zwei Konzepten.
Distanzbasierte Maße suchen in der Regel nach dem kürzesten Pfad zwischen zwei Konzepten. Bis auf die Arbeit von Sussna (1993), wird bei der Berechnung des kürzesten Pfades lediglich die *is-a* Relation berücksichtigt. Das Maß semantischer Verwandtschaft ist im einfachsten Fall die Pfadlänge selbst oder eine Funktion davon, vgl. z.B. Leacock & Chodorow (1998). Hirst & St-Onge (1998) gewichten die Pfade abhängig von ihrer Beschaffenheit, d.h. abhängig von verschiedenen Typen der im Pfad enthaltenen semantischen Relationen.
Die dritte Klasse der Methoden zur Bestimmung semantischer Verwandtschaft, informationsgehaltsbasierte Maße, benutzt die Struktur der GermaNet-Hierarchie und statistische Korpusauswertungen. Statistische Auswertungen dienen der Bestimmung des Informationsgehalts eines Konzeptes. Resnik (1995) definiert semantische Ähnlichkeit zwischen zwei Wörtern w_1 und w_2 als den maximalen Informationsgehalt ihres nächsten gemeinsamen Oberknoten *c* (Gleichung 2). c_1 und c_2 sind Konzepte (Wortbedeutungen), die w_1 und w_2 entsprechen. $S(c_1,c_2)$ ist eine Menge der Konzepte, die sowohl c_1 als auch c_2 subsumieren. $-\log p\,(c)$ ist der Informationsgehalt eines Konzeptes. Die Wahrscheinlichkeit *p* wird als die relative Frequenz der Wörter in einem Korpus berechnet. Lin (1998) definiert semantische Ähnlichkeit mit Hilfe des Informationsgehalts und eines Modells aus der Informationstheorie. Sein Maß (Gleichung 3) wird manchmal als universales Maß semantischer Ähnlichkeit genannt, da es anwendungs-, domänen-, und ressourcenunabhägig ist.

$$sim(c1,c2) = \max_{c \in S(c1,c2)} \left[-\log p(c) \right]$$

Gleichung 2

$$sim(c1,c2) = \frac{2 \times \log p(lcs(c1,c2))}{\log p(c1) + \log p(c2)}$$

Gleichung 3

Gurevych & Niederlich (2005b) führten eine komparativ ausgerichtete Evaluierung der Maße semantischer Verwandtschaft mit einem deutschen Datensatz (65 aus Substantiven bestehende Wortpaare) durch. Die Evaluierung ergab, dass wörterbuchbasierte und informationsgehaltsbasierte Maße semantischer Verwandtschaft

etwa gleich gut abschneiden. Wörterbuchbasierte Maße ermöglichen Vergleiche zwischen Wörtern unterschiedlicher Wortarten. Informationsgehaltsbasierte Maße erreichen zwar eine etwas bessere Performanz, der Vergleich ist bei diesen Maßen dafür immer innerhalb nur einer bestimmten Wortart möglich, da einzelne Wortarten in GermaNet in verschiedenen Taxonomien modelliert sind. Ebenso stellte sich als problematisch heraus, von unterschiedlichen Methoden berechnete Werte auf einen einheitlichen numerischen Bereich abzubilden, da die Verfahren unterschiedlicher Natur sind.

Da die Methode von Lin (1998) in der Evaluierung mit den deutschen Wortpaaren gut abgeschnitten hatte, wurde dieses Verfahren in das IR-Modell eingebunden. Die vom Verfahren berechneten Werte liegen auf der Skala von 0 bis 1. Die Repräsentation der Texte wurde auf eine Menge von Nomen begrenzt. Zusätzlich wird in der Vorverarbeitung eine domänenspezifische Stoppwortliste angewandt. Die Stoppwortliste bezieht sich auf die Berufsberatungsdomäne und enthält 137 Einträge, z.B. „Abschlussbezeichnung", „Ausbildungszeit". Dokumente und Anfragen werden als eine Menge der dort vorhandenen GermaNet-Konzepte dargestellt: *K_d* oder *K_q={k₁, ..., kₙ}*. Für jedes Paar *K_d* und *K_q* wird eine zweidimensionale Matrix *(#K_d x #K_q)* erstellt, in der # für die Anzahl der Elemente in der Menge steht. Dann wird für jedes Konzeptpaar ein Wert semantischer Verwandtschaft berechnet. Die Relevanz eines Dokuments *Rl(d,q)* zur Anfrage wurde nach zwei Gleichungen (4 und 5) ermittelt.

$$Rl(d,q) = \frac{\sum_{i=1}^{\#K\_d} \sum_{j=1}^{\#K\_q} rel(i,j)}{\#K\_d \times \#K\_q}$$

Gleichung 4

$$Rl(d,q) = \frac{\sum_{i=1}^{\#K\_d} \sum_{j=1}^{\#K\_q} rel(i,j)}{\#K\_q}$$

Gleichung 5

rel (i,j) steht für einen einzelnen Wert der semantischen Verwandtschaft zwischen zwei Konzepten. Da eine große Anzahl der Wortpaare einen geringen Wert der semantischen Verwandtschaft hat und im *Information Retrieval* Kontext nur eng verwandte Wortpaare von Bedeutung sind, wurde ein Schwellwert implementiert. Wortpaare mit einem Verwandtschaftswert unter diesem Schwellwert wurden bei der Berechnung des Gesamtwertes *Rl(d,q)* nicht berücksichtigt. Ergebnisse dieser Experimente sind in bezug auf verschiedene Schwellwerte und zwei Gleichungen in der Tabelle 2 dargestellt.

Schwellwert	Gleichung 3	Differenz	Gleichung 4	Differenz
kein	0,224	-0,0947	0,2813	-0,0374
0,3	0,2555	-0,0632	0,3035	-0,0152
0,5	0,2543	-0,0644	0,3114	-0,0073
0,7	0,2614	-0,0573	0,3129	-0,0058
0,8	0,281	-0,0377	0,3292	**0,0105**
0,9	0,2787	-0,04	0,3162	-0,0025
Baseline	0,3187			

Tabelle 2

Die Analyse der Ergebnisse macht deutlich, dass der auf semantischer Verwandtschaft basierende IR-Ansatz ungefähr gleich gut abschneidet, wie das konventionelle IR-System. Nur in einer Systemkonfiguration (Gleichung 5, Schwellwert 0,8) sind die Ergebnisse leicht besser als die Baseline. Es ist zu erkennen, dass der Schwellwert eine wichtige Rolle für die Endergebnisse spielt. So verbessern sich die Ergebnisse des semantischen IR-Systems mit steigendem Schwellwert. Das deutet darauf hin, dass die Berücksichtigung nur sehr eng verwandter Wörter, z.B. Synonyme oder direkte Hyponyme eine positive Auswirkung auf das Information Retrieval hat.
Eine detaillierte Auswertung der log-Dateien offenbarte eine Reihe der Fehlerquellen im System. Das semantische IR berücksichtigt nur Wörter, für die ein Verwandtschaftswert berechnet werden konnte. Einige Wörter, insbesondere viele zusammengesetzte Begriffe (Komposita) konnten nicht auf das GermaNet abgebildet werden, da Komposita eine offene Wortklasse im Deutschen sind und in einer statischen Wissensquelle wie GermaNet nicht im großen Umfang repräsentiert werden können. Folglich wurden Komposita bei der Berechnung semantischer Verwandtschaft nicht berücksichtigt.
Ähnlich wie in Experimenten mit Anfrageerweiterung führte das *Stemming* zu Fehlern in der Abbildung auf GermaNet. Daher entstehen manchmal sinnlose Verbindungen, z.B. bei „Blut – Schulzeit“ ergibt sich der Wert 0,85, da „Blut“ fehlerhaft auf „Blüte“ abgebildet wird und beide durch den Oberbegriff „Lebensphase“ miteinander verknüpft werden. Um derartigen Fehlern vorzubeugen, sollte das *Stemming* während der Abbildung auf GermaNet in der Anwendung vorzugsweise durch morphologische Analyse ersetzt werden. Des Weiteren scheint eine Repräsentation des Dokuments durch eine Menge der Nomen unzureichend (zu oberflächlich) zu sein. Wenn im Interessenprofil das Wort „Lehrer“ vorkommt und eine Berufsbeschreibung die Schulfächer des jeweiligen Ausbildungsberufes auflistet, werden die Schulfächer und „Lehrer“ fehlerhaft in einen engen Zusammenhang gebracht („Lehrer – Mathematik“ 0,756, „Lehrer – Biologie“ 0,739, „Lehrer – Seminare“ 0,70). Um diesen Effekt zu vermeiden, wäre es notwendig, semantische Struktur der Berufsbeschreibungen zu erkennen und Interessenprofile nur mit bestimmten Abschnitten, z.B. Aufgabenbeschreibungen des Berufes, zu vergleichen.
Eine weitere grundsätzliche Schwierigkeit (für das semantische IR und das Baseline-Verfahren) ist, dass die Anfragen nicht als präzise definierte Schlagwörter vorliegen, sondern als natürlichsprachlich formulierte Interessenprofile. Diese werden als *bag-of-words* aufgefasst und auf eine Menge der Nomen abgebildet. Verben, z.B. „Ich backe gerne“, oder Adverbien, z.B. „Ich arbeite gerne handwerklich“ werden nicht berücksichtigt. Negationen und verschiedene andere linguistische Mittel um negative Einstellungen einer Person wiederzugeben können zur Zeit nicht behandelt werden, was oft dazu führt, das unerwünschte Berufe zurückgeliefert werden. Ebenso wäre es wünschenswert, einzelne in der Anfrage vorkommende Begriffe nach ihrer Wichtigkeit für den Text zu gewichten. Dazu könnten fortgeschrittene Vorverarbeitungsverfahren, z.B. lexikalische Ketten, eingesetzt werden.

4. Diskussion

In diesem Beitrag wurden Möglichkeiten beschrieben, das semantische Wissen über Konzepte im *Information Retrieval* zu nutzen. Das zugrunde liegende lexikalisch-semantische Wissen wurde aus dem deutschen Wortnetz GermaNet bezogen. Experimentelle Arbeiten wurden an einem Testdatensatz in der Domäne „Elektronische Berufsberatung“ durchgeführt. Dabei wurde die Berufsberatung als eine *Information Retrieval* Aufgabe definiert.

Die Ergebnisse der semantisch angereicherten IR-Systeme wurden gegen einen automatisch erzeugten *Gold Standard* abgeglichen. Zum Vergleich wurde ein konventionelles IR-System herangezogen, das sich Zeichenkettenvergleiche und statistischer Wahrscheinlichkeiten bedient. Experimente mit automatisch erzeugten Definitionen der Konzepte haben ergeben, dass die Hyponymie-Relation sich als nützlich im Kontext des *Information Retrieval* erweist. Alle anderen semantischen Relationen haben sich dagegen als eher schädlich erwiesen. Es gilt zu überprüfen, ob das Hinzufügen einer Komponente zur Wortlesartendisambiguierung die Ergebnisse positiv beeinflussen kann.

Die semantische Verwandtschaft der Wörter wurde als Modell für die Bestimmung der Relevanz eines Dokuments eingesetzt. Dieses Modell hat eine ähnliche Performanz ermöglicht wie herkömmliche Ansätze. Um das System zu verbessern, müssen auf der einen Seite die Maße semantischer Verwandtschaft verbessert werden (z.B. im Hinblick auf die Behandlung mehrerer Wortarten). Auf der anderen Seite müssen zusätzliche Komponenten in der Vorverarbeitung eingesetzt werden (z.B. für die Kompositaanalyse, Wortartenerkennung). Die Aufbereitung der natürlichsprachlichen Interessenprofile als eine IR-Anfrage sollte verbessert werden, indem relevante Inhaltswörter sowie positive/negative Präferenzen eines Benutzers bestimmt werden.

Wichtig ist für zukünftige Arbeiten, die am Fallbeispiel „Elektronische Berufsberatung“ erzielten experimentellen Ergebnisse auf einem größeren und neutralen Datensatz zu verifizieren. Für die deutsche Sprache ist mit der sozialwissenschaftlichen Datenbank GIRT eine solche Möglichkeit gegeben (Kluck, 2004).[7] Die Stärken und Schwächen des semantischen *Information Retrievals* können dann in einem Vergleich mit anderen aktuellen Systemen besser verstanden werden.

Danksagung

Wir danken der Klaus Tschira Stiftung für die finanzielle Unterstützung und der Bundesagentur für Arbeit für die zur Verfügung gestellten Daten. Hendrik Niederlich möchte ich besonders für seine wertvollen Beiträge zu dieser Studie danken, die er als Praktikant und Diplomand am EML Research geleistet hat.

[7] http://www.gesis.org/Forschung/Informationstechnologie/GIRT4.htm

Bibliographie

Martha Evens. 2002. Thesaural relationships in information retrieval. In Rebecca Green, Carol Bean, and Sung Hyon Myaeng, editors, *The Semantics of Relationships. An Interdisciplinary Perspective*, Chapter 9. Kluwer Academic Publishers, Dordrecht.

Julio Gonzalo, Felisa Verdejo, Irina Chugur, and Juan Cigarran. 1998. Indexing with WordNet synsets can improve text retrieval. In *Proceedings of the COLING-ACL '98 Workshop Usage of WordNet in Natural Language Processing Systems,* Montreal, Canada, August.

Iryna Gurevych und Hendrik Niederlich. 2005a. Computing Semantic Relatedness in German with Revised Information Content Metrics. In *Proceedings of "OntoLex 2005 – Ontologies and Lexical Resources" IJCNLP'05 Workshop*, Jeju Island, Republic of Korea, October 15, 2005. *To appear.*

Iryna Gurevych und Hendrik Niederlich. 2005b. Computing semantic relatedness of GermaNet concepts. In Bernhard Fisseni, Hans-Christian Schmitz, Bernhard Schröder, and Petra Wagner, editors, *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Applications of GermaNet II*, pages 462–474. Peter Lang.

Iryna Gurevych. 2005a. Automatically generating a task-based information retrieval test collection. Technical Report, EML Research, Heidelberg, 2005.

Iryna Gurevych. 2005b. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP'2005),* Jeju Island, Republic of Korea, October 11–13, 2005. *to appear.*

Graeme Hirst und David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database and some of its applications*, pages 305–332. Cambridge, MA: The MIT Press.

Michael Kluck. 2004. Die GIRT-Testdatenbank als Gegenstand informationswissenschaftlicher Evaluation. Bernard Bekavac, Josef Herget, Marc Rittberger, editors, Informationen zwischen Kultur und Marktwirtschaft. *Proceedings des 9. InternationalenSymposiums fur Informationswissenschaft (ISI 2004)*, Chur, 6.-8.Oktober 2004. Konstanz: UVK Verlagsgesellschaft mbH, 2004. S. 247 – 268.

Claudia Kunze. 2004. Lexikalisch-semantische Wortnetze. In K.-U. Carstensen, C. Ebert, C. Endriss, S. Jekat, R. Klabunde, and H. Langer, editors, *Computerlinguistik und Sprachtechnologie. Eine Einführung*, pages 423–431. Heidelberg, Germany: Spektrum Akademischer Verlag, 2nd edition.

Claudia Leacock und Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 265–283. Cambridge: MIT Press.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation,* Toronto, Ontario, Canada, June, pages 24–26.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning,* San Francisco, Cal., pages 296–304.

Rila Mandala, Takenobu Tokunaga, Hozumi Tanaka, Akitoshi Okumura, und Kenji Satoh. 1998. Ad hoc retrieval experiments using WordNet and automatically constructed thesauri. In *Text REtrieval Conference*, pages 414–419.

Phil Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence,* Montreal, Canada, 20–25 August 1995, volume 1, pages 448–453.

Mark Sanderson. 1994. Word sense disambiguation and information retrieval. In SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pages 142 - 151, Springer-Verlag New York, Inc., New York, NY, USA.

G. Salton, E. Fox, und H. Wu. 1983. Extended boolean information retrieval. Communications of the ACM, 26(11): 1022-1036.

Alan F. Smeaton. 1999. Using NLP or NLP resources for information retrieval tasks. In Tomek Strzalkowski, editor, *Natural language information retrieval*, pages 99–111. Kluwer Academic Publishers, Dordrecht, NL.

Michael Sussna. 1993. Word sense disambiguation for free text indexing using a massive semantic network. In Proceedings of the 2nd International Conference on Information and Knowledge Management (CIKM'93), Arlington, Virginia.

Ellen M. Voorhees. 1994. Query expansion using lexical-semantic relations. In *SIGIR'94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69, New York, NY, USA. Springer-Verlag New York, Inc.

Ellen M. Voorhees. 1999. Natural language processing and information retrieval. In *Information Extraction: Towards Scalable, Adaptable Systems*, pages 32–48, London, UK. Springer-Verlag.