Lernen paradigmatischer Relationen auf iterierten Kollokationen

Christian Biemann, Stefan Bordag, Uwe Quasthoff
Universität Leipzig
Augustusplatz 10/11
04109 Leipzig
{biem, sbordag, quasthoff}@informatik.uni-leipzig.de

Abstract

Das Lernen paradigmatischer Relationen wie Synonymie, Homonymie, Antonymie und Hyponymie ist Thema verschiedener statistischer Ansätze. Die bisherigen Ansätze verwenden nur je ein statistisches Feature, um derartige Relationen aus großen Textkorpora zu extrahieren. In diesem Papier soll eine Architektur vorgestellt, die es ermöglicht, Relationen zwischen Wörtern durch eine Trainingsmenge zu lernen, um weitere in der Relation stehende Wörter zu erhalten, um schließlich lexikalisch-semantische Wortnetze automatisch oder halbautomatisch zu erweitern. Hierzu wird zunächst eine passende Menge von Features aus einer großen Menge vorhandener Features aufgrund der Trainingsmenge ausgewählt, statistisch getestet und zum Erweitern des Wortnetzes verwendet.

1 Einleitung

Ein in den letzten Jahren wiederholt angesprochenes Problem der linguistischen Datenverarbeitung ist das „acquisition bottleneck“: Die meiste Zeit wird darauf verwendet, lexikalische Ressourcen manuell aufzubauen. Wird auch die Wiederverwendung durch Standards wie MLEXd [siehe Ahmad 1994] oder TEI [Sperberg-McQueen et al. 02] unterstützt, so benötigt dennoch beinahe jede neue Anwendung andere Angaben in den jeweiligen Lexika oder Wortnetzen.

Wir wollen einen Ansatz vorstellen, wie vorhandene Wortnetze automatisch bzw. semiautomatisch unter Zuhilfenahme großer Korpora erweitert werden können.

Eine Teilaufgabe bei der Erweiterung von Ressourcen wie GermaNet (http://www.sfs.uni-tuebingen.de/lsd/Intro.html) ist das Zuordnen von bisher nicht aufgenommenen Wörtern zu vorhandenen Synsets oder Gruppen von Synsets.

Vorhandene Verfahren zum Auffinden von Synonymen oder Pseudosynonymen in großen Dokumentenkollektionen wie [Ruge 97] oder [Rapp 02] nehmen an, dass ähnliche Wörter in ähnlichen Umgebungen zu finden sind und benutzen als Umgebung einerseits Dependenzen, andererseits das gemeinsame Vorkommen in Sätzen.

Nach Eingabe eines Wortes liefern derartige Verfahren eine gerankte Liste von ähnlichen Wörtern. Die Qualität ist jedoch für unsere Zwecke nicht ausreichend, da zwar Synonyme im Allgemeinen gefunden werden, aber auch viele anderweitig verwandte Wörter.

Unser Ansatz versucht, durch den Vergleich mehrerer derartiger Listen die Synonyme von den übrigen Wörtern zu trennen.

Der Ansatz ähnelt dem in [Chiaramita 02], wo versucht wird, mit Hilfe von morphologischen Merkmalen Wörter zu weit gefassten Wordnet-Synsets zuzuordnen, benutzt jedoch mehr Features.

2 Verfahren

Unsere Aufgabe ist das Erweitern von Wortmengen, z.B. Synsets aus GermaNet. Das Vorgehen gliedert sich in zwei Schritte: Im ersten Schritt werden mittels verschiedener Verfahren weitere Wortmengen erzeugt, deren Elemente sich möglicherweise zur Erweiterung einer vorgegebenen Menge eignen. Da möglicherweise diese Mengen von unterschiedlicher Qualität sein können, werden im zweiten Schritt daraus geeignete Mengen selektiert und aus den Mengen geeignete Elemente zur Erweiterung extrahiert.

2.1 Mengen konstruieren

Die uns interessierenden Synsets bestehen in der Regel aus (Quasi-)Synonymen. D.h., die enthaltenen Wörter stimmen in vielen wichtigen semantischen Eigenschaften überein, unterscheiden sich möglicherweise aber auch in einem Attribut.

Zur Konstruktion der im zweiten Schritt zur Erweiterung verwendeten Mengen benutzen wir nun Verfahren, die korpusbasiert möglichst semantisch homogene Wortmengen liefern. Semantische Homogenität soll hier bedeuten, das sich die Elemente einer solchen Menge möglichst einfach beschreiben lassen, beispielsweise durch Variation nur eines semantischen Attributes.

Die automatische Konstruktion solcher Mengen ist selbstverständlich schwierig, in der Regel entstehen Mengen, die diese Eigenschaft nur näherungsweise erfüllen. Bei vielen Verfahren entsteht zusätzlich ein Ranking, d.h. ein numerischer Wert, welcher die Rangordnung in der Menge beschreibt. Hier kann mit einem Schwellwert Einfluss auf die Größe der erzeugten Mengen genommen werden. Mit einem hohen Schwellwert kann möglicherweise eine höhere Qualität der erzeugten Mengen gesichert werden.

Im Abschnitt 2 werden mehrere Verfahren zur Erzeugung solcher Mengen vorgestellt.

2.2 Synsets erweitern

Im zweiten Schritt sollen aus unserem eben erzeugten Reservoir an semantisch möglichst homogenen Mengen diejenigen ausgewählt werden, die sich zur Erweiterung eines gegebenen Synsets eignen. Das Problem besteht darin, dass zwar auch das Synset semantisch homogen ist (oder wenigstens sein sollte), aber die semantische Homogenität kann sich auf ein anderes Attribut beziehen. Zur Erweiterung eignen sich also nur solche Mengen, die bezüglich desselben Attributs semantisch homogen sind wie das vorgegebene Synset. Da weder die semantischen Attribute noch ihre Werte bekannt sind, muss die Übereinstimmung an Hand gleicher Elemente getestet werden: Zwei semantisch homogene Mengen haben diese Eigenschaft bezüglich desselben Attributs, falls sie möglichst viele Elemente gemeinsam enthalten.

Unser Vorgehen gestaltet sich damit folgendermaßen: Um ein gegebenes Synset $S=\{s_1, ..., s_n\}$ zu erweitern, suchen wir in unserem Reservoir an semantisch homogenen Mengen eine Menge $M=\{m_1, ..., m_k\}$ mit möglichst großem Durchschnitt $S \cap M$. Die zusätzlichen Elemente $M \setminus S$ sind die Kandidaten für die Erweiterung von S.

Falls sich die erzeugte Erweiterungsmenge M als nicht ausreichend semantisch homogen erweist, lassen sich alternativ zunächst mehrere Mengen $M_1, ..., M_r$ auswählen und zur tatsächlichen Erweiterung nur die Durchschnittsmenge $M=M_1 \cap ... \cap M_r$ benutzen.

Die Schnittmengenoperation kann an dieser Stelle auch derartig abgeschwächt werden, dass ein Wort auch dann in der Ergebnismenge zu finden ist, wenn es in den meisten Mengen M_i mit hohem Rang vorkommt.

Für das Verfahren sind an zwei Stellen ausreichende Datenmengen nötig: zum einen ein ausreichend großes Korpus, um nicht am *data sparseness problem* für statistische Verfahren zu scheitern, zum anderen eine genügend große zu erweiternde Wortmenge, um die Art der Homogenität automatisch erfassen zu können.

Letzteres Problem lässt sich für GermaNet derart behandeln, dass nicht z.B. Synsets als Wortmenge erweitert werden, sondern Vereinigungen von Synsets, die in der Hierarchie benachbart liegen.

Große Korpora stehen bereit, siehe dazu Abschnitt 2.3.1.

2.3 Wortmengen verschiedener Eigenschaften

2.3.1 Kollokationen

Seit 1993 wird im Rahmen des Projekts *Deutscher Wortschatz* eine umfangreiche Textsammlung [siehe Quasthoff 1998] gepflegt, für welche u. a. statistische Kollokationen auf Satzbasis und Nachbarschaftsbasis berechnet werden, siehe dazu [Heyer et al. 01].

Zu jedem Wort kann eine nach Signifikanz geordnete Liste von Wörtern extrahiert werden, die mit diesem Wort statistisch auffällig gemeinsam auftreten. Dieses gemeinsame Auftreten wird dabei unterschieden nach unmittelbaren linken und rechten Nachbarn, sowie dem Auftreten im gleichen Satz.

Durch solche Kollokationen werden zwar menschliche Assoziationen reflektiert, also semantische Zusammenhänge ausgedrückt, aber in der Regel ist eine solche Kollokationsmenge seman-

tisch recht inhomogen, weil verschiedenartige Assoziationen möglich sind. Diese Inhomogenität wird bei einer möglichen Polysemie des Ausgangswortes natürlich noch verstärkt.

Aus diesen Gründen sind reine Kollokationsmengen für die Erweiterung von Synsets zunächst ungeeignet, wie die Kollokationsmenge für „Witterung“ aufzeigt (Die Zahlen in Klammern geben Signifikanzen an):

Kollokationen für Witterung:

kühler (115), ungünstiger (85), schlechter (81), milde (77), kühle (72), kalte (69), milden (69), kühlen (66), wegen (62), feuchte (57), schlechten (48), warme (43), naßkalten (40), naßkalter (39), günstiger (38), warmen (38), kalten (37), aufgenommen (36), feuchter (35), kalter (35), trotz (35), Unbilden (34), ungünstigen (29), warmer (28), ungünstige (27), Wegen (26), aufgrund (24), günstige (24), Bei (22), Feldberg (22), anhaltend (22), zuläßt (21), Jahreszeit (20), Temperaturen (20), Ts (20), ausgesetzt (20), naßkalte (20), ungewöhnlich (19), je (18), schlechte (17), trockener (17), Sommer (16), angenehmer (16), günstigen (16), Regen (15), geschützt (15), wechselhafte (15), Heizöl (14), feuchten (14), vergangenen (14), Begünstigt (13), Grad (13), besonders (13), Pilz (12), Trauben (12), begünstigt (12), verlegt (12), widrigen (12), Baaderplatzes (11), Bauarbeiten (11), Bäume (11), Heizgradtage (11), Meteorologen (11), Sobald (11), Trotz (11), Volksheilkundliche (11), extrem (11), kühlere (11), nasse (11), selbstgefressenen (11), waldfreundliche (11), widriger (11), winterlichen (11), Freien (10), Schnee (10), Schneekappe (10), Tharaus (10), Warenhaus-Manager (10), Wetter (10), Wildkräuterführungen (10)

2.3.2 Filtern von Kollokationsmengen

In vielen Fällen drückt sich die eben beschriebene semantische Inhomogenität einer Kollokationsmenge auch durch die Wortart der entsprechenden Wörter aus. Eine einfache Filterung nach Wortarten erzeugt häufig gute Mengen zur Weiterverarbeitung. Solche Filterkriterien sind beispielsweise:

- Für Adjektive: spezielle Substantive als rechte Nachbarn
- Für Substantive: spezielle Adjektive als linke Nachbarn
- Für Substantive: spezielle Verben als linke oder rechte Nachbarn
- Für Substantive: Substantive, die Satzkollokationen, aber keine Nachbarschaftskollokationen sind.

2.3.3 Iterierte Kollokationen

Aus diesen Kollokationen erster Stufe können iterativ Kollokationen höherer Stufe erzeugt werden: Für die Kollokationen zweiter Stufe wird ein Korpus benutzt, welches aus Kollokationsmengen von Kollokationen erster Stufe besteht. Diese Kollokationsmengen übernehmen die Rolle der Sätze des Korpus. Da die Reihenfolge der Wörter in den Kollokationsmengen nicht aussagekräftig ist, wohl aber ihr gemeinsames Vorkommen, sind hier nur Satzkollokationen sinnvoll. In der dritten Stufe werden statt Sätzen die Kollokationsmengen zweiter Stufe ausgewertet und so weiter.

Hohe Signifikanzen zweier Wörter in den Kollokationen zweiter Stufe bedeuten, dass diese Wörter häufig gemeinsam in Kollokationsmengen erster Stufe auftreten. Dies wiederum bedeutet, dass die entsprechenden Wörtern in vielen Kontexten gemeinsam auftreten und damit zu einem größeren Kontext gehören. Diese Eigenschaft erinnert sofort an die Zusammenfassung von Wörtern entsprechend Bedeutungsgruppen, oder andersherum, an die Zerlegung von Kollokationsmengen erster Ordnung entsprechend verschiedener Bedeutungen.

Die inhaltliche Bedeutung von Kollokationen höherer Ordnung ist zunächst nicht klar. Aus den Berechnungen bis Stufe 10 (um evtl. reine Mengen zu erhalten) ist beobachtbar, dass die Kollokate höherer Stufe für viele Wörter semantisch relativ homogene Wortmengen darstellen. Allerdings ist nicht unbedingt vorhersehbar, auf welches semantische Attribut sich die Homogenität bezieht. Dies kann abhängen von der Art der verwendeten Kollokationsmengen (Satzkollokationen oder Nachbarschaftskollokationen) im ersten Schritt, ebenso von der Größe der verwendeten Kollokationsmengen. Hier scheinen größere Kollokationsmengen allgemeinere Zusammenhänge zu extrahieren.

Selbstverständlich können auch diese iterierten Kollokationsmengen mit Filtern wie aus Abschnitt 2.3.2. behandelt werden.

2.4 Beispiele

Im verbleibenden Teil wird anhand zweier Beispiele illustriert, wie Kohyponyme mit Kookkurrenzen verschiedener Stufen automatisch extrahiert werden können.

2.4.1 Kohyponyme über Nachbarschaftskollokationen zweiter Stufe

Als Beispiel betrachten wir die Kohyponyme "warm", "kalt" und "kühl". Zu diesen Wörtern berechnen wir die Nachbarschaftskollokationen zweiter Stufe, also Wörter, die in einem größeren Kontext zusammen mit den Startwörtern auftreten.. Ferner filtern wir die drei Wortmengen, so dass nur Adjektive übrigbleiben. Tabelle 1 zeigt einen Ausschnitt aus den drei alphabetisch sortierten Wortmengen..

warm	**kühl**	**kalt**
abgekühlt	abgeklärt	abgekühlt
abkühlen	abgekühlt	abkühlen
angestiegen	abkühlen	angestiegen
anzeigt	ablehnend	anzeigt
aufgeheizt	abstrakt	aufgeheizt
eingefroren	aggressiv	aushalten
erhitzt	ähnlich	eingefroren
erwärmt	altmodisch	einstellen
fertig	anders	erhitzt
gebrannt	archaisch	ernst
gefallen	aufgeheizt	erwärmt
gehalten	aushalten	frei
geklettert	bedrohlich	gebrannt
gekühlt	bescheiden	gefallen
gelagert	bitter	gehalten
gemessen	blaß	geklettert
gesenkt	blutleer	gekühlt
gestiegen	distanziert	gelagert
gesunken	eingefroren	gemessen
gut	empfindlich	genug
Heiß	empört	gesenkt
heruntergekühlt	entrüstet	gestiegen
hoch	entsetzt	hart
höher	entspannt	heiß
kalt	erhitzt	heruntergekühlt
kalte	erleichtert	hoch
kalten	erschöpft	höher
...	...	...

Tabelle 1: Nachbarschaftskollokationen zweiter Stufe mit Adjektivfilter für „warm“, „kühl“ und „kalt“.

Die Schnittmenge dieser drei Mengen enthält die Wörter *abgekühlt, aufgeheizt, eingefroren, erhitzt, erwärmt, gebrannt, gelagert, heiß, heruntergekühlt, verbrannt* und *wärmer*, also bis auf ein Wort Abstufungen von Temperatur. Die emotionale Lesart von „kalt“ und „kühl“, die sich in Wörtern wie *abgeklärt* ablesen lässt, wird vollständig durch den Schnitt mit der Menge zu „warm“ eliminiert.

2.4.2 Kohyponyme mit Kollokationen erster Stufe

In diesem Abschnitt wird eine weitere Möglichkeit gezeigt, semantische Relationen zu extrahieren, und zwar insbesondere linguistische Kollokationen, Kohyponyme oder Hyperonyme, wobei hier zwischen diesen Arten unterschieden werden kann. Diese Möglichkeit ergibt sich aus der Parallele zwischen syntagmatischen sowie paradigmatischen Beziehungen de Saussures, siehe [Saussure 1916], und der Mengen von Satzkollokationen einzelner Wortformen. Während gemeinsames Auftreten in einem Satz grob als syntagmatische Relation eingeschätzt werden kann, soll im Folgenden eine paradigmatische Relation beschrieben werden. Zu jedem Wort einer Menge von Satzkollokationen werden wiederum dessen Satzkollokationen ermittelt. Diese Mengen werden verglichen, indem der Anteil gemeinsam auftretender Elemente bestimmt wird. Damit haben wir für zwei Wörter A und B zwei Zahlenwerte ermittelt: Die Kollokationsstärke zwischen A und B sowie die Ähnlichkeit der Kollokationsmengen von A und von B. Trägt man für ein festes Wort A alle seine Kollokate entsprechend dieser beiden Werte in einem Koordinatensystem ab, so beobachtet man folgendes: Kohyponyme weisen einen hohen direkten Signifikanzwert auf und besitzen auch ähnliche Kollokationsmengenund befinden sich damit im rechten oberen Bereich der graphischen Darstellung. Hyperonyme hingegen dürften selten im gleichen Satz auftreten, ansonsten aber ein sehr vergleichbares Kollokationsprofil aufweisen und damit im äußersten unteren Bereich der Abbildung zu finden sein. Linguistische Kollokationen hingegen müssten geradezu umgekehrt, zwar häufig im gleichen Satz auftretend, ansonsten aber sehr unterschiedliche Kollokationprofile besitzen und sich damit im linken oberen Bereich der Abbildung aufhalten.

Eine prototypische Implementierung dieses Verfahrens hat gezeigt, dass es wie erwartet die drei genannten Relationen extrahiert. Es hat aber auch

gezeigt, dass es Wörter gibt, für die fälschlicherweise Kohyponyme extrahiert werden, obwohl diese keine besitzen. Es wird demnach noch ein weiteres Feature benötigt, um die richtigen Wörter von den unpassenden unterscheiden zu können. Weiterhin ist die auf diese Weise erhaltene Kohyponymmenge oft durch in der ‚part-of'-Relation stehende Wörter verunreinigt. Schließlich aber befinden sich in der Nähe der erwarteten Hyperonyme oft noch andere, unbeteiligte Wörter, wofür ebenfalls ein anderes Filterkriterium gefunden werden muss.

Es sei jedoch angemerkt, dass es nicht wichtig ist, die Grenzen zwischen den einzelnen räumlich getrennten Bereichen für eine optimale Zuordnung zu kennen, zudem sich diese Grenzen von Wortform zu Wortform leicht unterscheiden können. Da jedoch eine Wortform mehrere Kohyponyme hat und idealerweise lediglich ein Hyperonym, können sich die Ergebnisse der Berechnungen der Kohyponyme gegenseitig verifizieren und so zu einem meist eindeutigen Ergebnis kommen.

Abbildung 1 zeigt ein Beispiel anhand „Elefant":

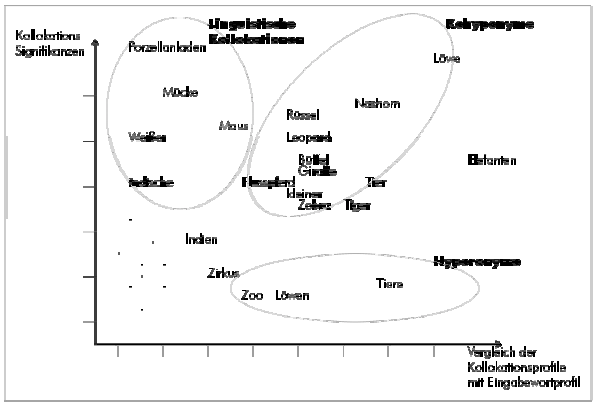

Abbildung 1: Koordinatensystem für Elefant mit Zuweisung von linguistischen Kollokationen, Hyperonymen und Kohyponymen

3 Implementierung und Evaluation

In diesem Abschnitt stellen wir Ergebnisse, die wir mit einer prototypischen Implementierung eines Verfahrens wie in 2.2. beschrieben erzielen konnten.

Zum Einsatz kamen Nachbarschaftskollokationen zweiter und dritter Stufe, sowie ein Wortartenfilter.

Intuitiv enthält die Nachbarschaftskollokationsmenge zweiter Stufe eines Wortes diejenigen Wörter, die ähnliche Bigrammkontexte aufweisen, der Wortartenfilter stellt sicher, dass nur Wörter mit derselben Wortart wie die Wörter der zu erweiternden Menge ausgegeben werden.

Zum Testen verwendeten wir einerseits Synsets aus drei verschiedenen lexikalischen Feldern aus GermaNet, andererseits Synonyme aus unserer Wortschatz-Sammlung.

Evaluiert wurde, wie viele Synonyme und wie viele ‚Xonyme' (dies soll hier verwendet werden als Sammelbezeichnung für Antonyme, Hyponyme, Hyperonyme und Kohyponyme) unter den ersten 5 bzw. 10 ausgegebenen Wörtern zu finden waren. Die Reihenfolge der Ausgabe erfolgte nach Signifikanz der Kollokation wie in [Quasthoff 2002] beschrieben.

Getestet wurde mit Synset-Mengen der Mächtigkeit 1 und 2.

Tabelle 2 zeigt Synsets, die Top 10 der automatisch erzeugten Ergebnismenge, sowie deren Klassifizierung in Synonyme (S) und ander Xonyme (X) .

Synset	Ergebnismenge
erstklassig, prima	besser(X), exzellent(S), glänzend(S), gut(X)
Mörder, Killer	Täter(X), Verbrecher(X), Attentäter(X), Kriegsverbrecher(X), Straftäter(X), Räuber(X), Mann, Einbrecher(X), Terrorist(X), Brandstifter(X)
verrückt	gerne, enttäuscht, bekannt, albern(S), begeistert, bescheuert(S), erschöpft, frustriert, dünn
festigen, stabilisieren	sichern(S), sorgen, auswirken, beitragen(X), gefährden, stützen(S), erwarten, fördern(X), profitieren, bringen, führen
dunkelbraun	braun(X), hellbraun(X), rotbraun(X), orange(X), rot(X), rosa(X), violett(X), religiös, blutrot(X), graubraun(X)

Tabelle 2: Synsets und deren automatische Erweiterungen. Synonyme sind mit (S), Xonyme mit (X) gekennzeichnet.

Die Abbildungen 2a – 2b zeigen die Anteile von Synonymen, Xonymen und nicht in diesem Sinne verwandten Wörtern für verschiedene Arten von Startmengen, die Werte sind jeweils gemittelt für 10 zufällig ausgewählte Synsets.

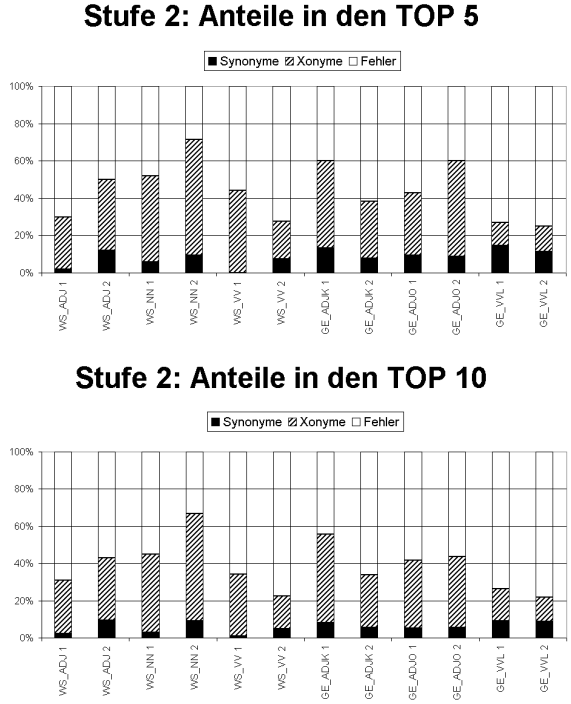

Abbildung 2a: Kollokationen zweiter Stufe

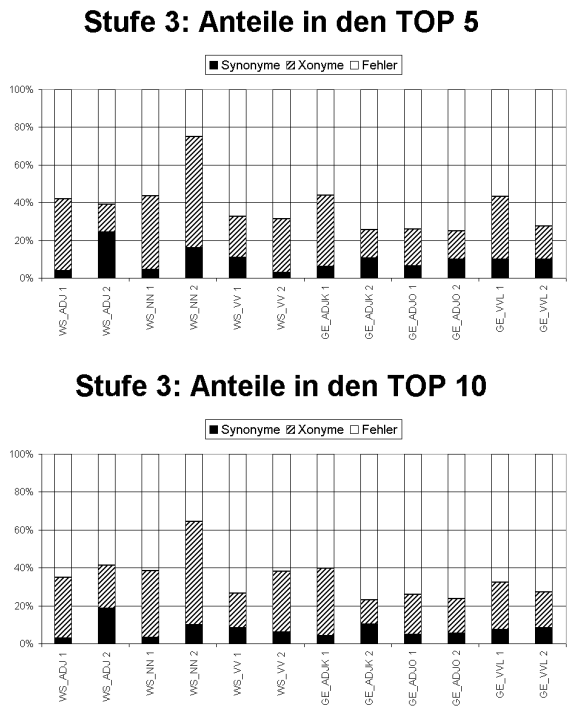

Abbildung 2b: Kollokationen dritter Stufe

Folgende Tabelle erklärt die in der Abbildung verwendeten Kürzel:

Kürzel	**Bedeutung**
WS_ADJ	Adjektive aus dem Wortschatz
WS_NN	Nomen aus dem Wortschatz
WS_VV	Verben aus dem Wortschatz
GER_ADJK	Adjektive aus GermaNet/Körper
GER_ADJO	Adjektive aus GermaNet/Ort
GER_VVL	Verben aus GermaNet/Lokation

Tabelle 3: Erklärung der Kürzel. Eine „1“ bzw. „2“ am Ende bedeutet, das das zu erweiternde Synset aus einem bzw. zwei Elementen bestand.

Der Synonym- und Xonymanteil in den ersten 5 Rängen ist in den meisten Fällen höher als in den ersten 10, was die Wirksamkeit unserer Ranking-Methode bestätigt.

Unsere Erwartungen, dass größere Startmengen bessere Ergebnisse bringen sollten, erfüllten sich nur teilweise: Während bei den Mengen aus dem Wortschatz die Qualität bei den Zweiermengen deutlich höher liegt als bei den Einermengen, zeigt sich bei den Germanet-Synsets ein ausgeglichenes Bild.

Insgesamt scheinen die Verben für nachbarschaftskontextbasierte Features wie die hier evaluierten am schlechtesten geeignet zu sein, was in Anbetracht ihrer komplexeren Argumentstruktur nicht verwunderlich ist. Die unterschiedliche Qualität bei den verschiedenen Wortarten motiviert das Lernen von Featurekombinationen wie in 2.2 beschrieben.

In der vorliegenden Form können die erzielten Daten lediglich als Vorschläge dienen, um GermaNet zu erweitern. Die endgültige Entscheidung über die Einordnung an der vorgeschlagenen Stelle muss manuell vorgenommen werden. Trotzdem entsteht bei dieser Methode durch das Vorschlagen eines Synsets schon eine Zeitersparnis von ca. 90% gegenüber der reinen Handarbeit.

4 Ergebnis und Ausblicke

Wir haben Möglichkeiten vorgestellt, korpusbasierte Erweiterungen von beliebigen Wortmengen durchzuführen.

Bei der Frage auf Anwendbarkeit auf semantische Netze wie GermaNet, muss nun folgendes beachtet werden: hat das zur Erweiterung vorgelegte semantische Netz eine gewisse Größe erreicht, so geht es bei der Einordnung nur noch um

mittel- und niederfrequente Wörter. Die Verfahren zur Erzeugung der Kandidatenmengen müssen also auch mit solchen Wörtern umgehen können.

Um zu verdeutlichen, dass das verwendete Korpus (36 Millionen deutsche Sätze) hierfür ausreichend ist, zeigt Abbildung 3 die durchschnittliche Größe der Kollokationsmengen für absteigend nach Frequenz geordnete Wörter (Vollformen).

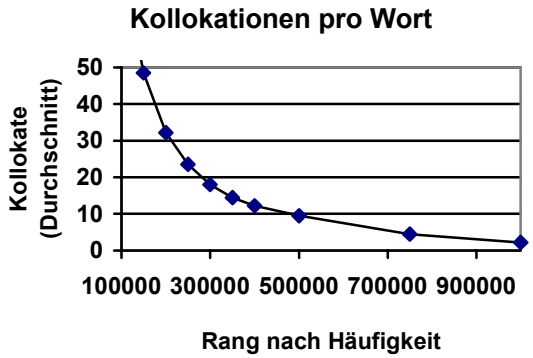

Abbildung 3: Durchschnittliche Anzahl Kollokate nach Häufigkeitsrang.

Um sinnvolle Ergebnisse zu erhalten, benötigen wir mindestens 10 Kollokate pro Wort, was bis Rang 500'000 der Fall ist. Unter der Berücksichtigung der Größe von GermaNet (etwa 61'000 Lexical Units) und der Struktur (es werden Lemmas gespeichert, während das Korpus mit Vollformen umgeht) erwarten wir, GermaNet in etwa um die Hälfte erweitern zu können.

Um auch niederfrequente Wörter extrahieren zu können, für die zu wenig Kollokate existieren, ist es nötig, auf musterbasierte Verfahren wie [Biemann 2003] zurückzugreifen, die auf Beispielsätzen arbeiten und auch mit Wortmengen trainiert werden können.

References

[Ahmad 94] Ahmad, K. (ed) (1994) MULTILEX: Final report. Guildford: University of Surrey

[Biemann 03] Biemann, C. (2003) Extraktion von semantischen Relationen aus natürlichsprachlichem Text mit Hilfe von maschinellem Lernen, In: Uta Seewald-Heeg (Hrsg.): Sprachtechnologie für multilinguale Kommunikation, Beiträge der GLDV-Frühjahrstagung 2003, Gardez! Verlag, Sankt Augustin 2003

[Bordag 2002] Stefan Bordag, Sentence Co-occurrences as Small-World Graphs: A solution to Automatic Lexical Disambiguation, A. Gelbukh (Ed.): CICLing 2003, LNCS 2588, pp. 329-332, Springer-Verlag Berlin Heidelberg, 2003

[Chiaramita 02] Chiaramita, M. (2002): Boosting automatic lexical acquisition with morphological information, Proceedings of Siglex02.

[Heyer et al. 01] Heyer, G.; Läuter, M.;Quasthoff, U.; Wittig, Th.; Wolff, Chr. (2001): Learning Relations using Collocations, In: A. Maedche, S. Staab, C. Nedellec and E. Hovy, (eds.). Proc. IJCAI Workshop on Ontology Learning, Seattle/ WA, 19. - 24. August 2001

[Quasthoff 98] Quasthoff, U. (1998): Projekt Der Deutsche Wortschatz . In Heyer, G.; Wolff, Ch. (eds.) (1998). Linguistik und neue Medien, Tagungsband zur GLDV-Tagung, 17. – 19. März 1997 in Leipzig, Deutscher Universitätsverlag, 93 - 99, 1998.

[Quasthoff 2002] Quasthoff, U.; Wolff, C. (2002): The Poisson Collocation Measure and its Applications. In: Proc. Second International Workshop on Computational Approaches to Collocations, Wien.

[Rapp 02] Rapp, R. (2002): The Computation of Word associations: Comparing syntagmatic and Paradigmatic Approaches, Proceedings of COLING-02, Taipei, Taiwan

[Ruge 97] Ruge, G., (1997): Automatic detection of Thesaurus Relations for Information Retrieval Applications, Lecture Notes on Computer Science, Vol. 1377

[Saussure 1916] Saussure, F de. (1916) : Cours de Linguistique Générale, Paris, Payot

[Sperberg-McQueen et al. 02] Sperberg-McQueen, C.M.. and Burnard, L. (eds.) (2002). TEI P4: Guidelines for Electronic Text Encoding and Interchange. Text Encoding Initiative Consortium. XML Version: Oxford, Providence, Charlottesville, Bergen).