Relationsfreie Kontextmodelle auf der Basis von *common sense* als integrierter Ansatz zur *word sense disambiguation* und Bildung lexikalischer Ketten

Christian Behrenberg

Ruhr-Universität Bochum
christian@behrenberg.de

Zusammenfassung

In dieser Arbeit wird ein Ansatz skizziert, mit dem Nomina anhand ihres lokalen Kontextes disambiguiert und zu lexikalischen Ketten verknüpft werden können. Dazu werden für Lemmata relationsfreie Kontextmodelle aufgestellt, für deren Konstruktion allgemeines Weltwissen (*common sense*) verwendet wird. Der Ansatz ist zur Umsetzung mit *WordNet* und *ConceptNet* als Wissensressourcen vorgesehen und wird stichprobenweise mithilfe des *SemCor*-Korpus evaluiert.

1. Einleitung

Lexikalische Ketten stellen eine Methode zur Identifizierung bedeutungsträchtiger Textfragmente dar. Sie können zur Lösung von Problemstellungen in *Natural Language Processing* (NLP) Anwendungen eingesetzt werden, wie z.B. zur automatischen Textzusammenfassung.

Die *word sense disambiguation* (WSD) ist dabei ein zugrundeliegendes Teilproblem und steht neben der Verbesserung der Laufzeit im Fokus der Forschungen: es stellt sich heraus, dass die Separierung der WSD vom eigentlichen *lexical chaining* nützlich ist. Allen relevanten Verfahren liegt dabei die Suche nach in Relation stehenden Wörtern zugrunde. Es lässt sich beobachten, dass die Kriterien, nach denen in Beziehung stehende Wörter im Text gesucht und bewertet werden, zwischen den Verfahren quantitativ und strukturell variieren; es ist unklar, inwieweit sie die WSD und das *lexical chaining* qualitativ beeinflussen.

Allen relevanten Verfahren liegt zudem die Wissensressource *WordNet* (siehe 2.4.1.) zugrunde. Da *WordNet* einerseits keine nicht-systematischen semantischen Relationen abbildet und andererseits Wörtern weitaus mehr Wortbedeutungen zuordnet, als man intuitiv annehmen würde, ist es fraglich, inwieweit *WordNet* als Grundlage für lexikalische Ketten geeignet ist, wenn man mit lexikalischen Ketten das Ziel verfolgt, allgemeine Themen in von Menschen geschriebenen Text zu registrieren und zu verknüpfen.

Die Motivation dieser Arbeit ist die Behandlung der geschilderten Probleme unter Berücksichtigung des aktuellen Forschungsstandes. Es wird im Folgenden ein Ansatz zur Bildung lexikalischer Ketten skizziert. Im Fokus der Arbeit steht die Verwendung von allgemeinen Weltwissen (*common sense*) zur Aufstellung von Kontextmodellen, die es gestatten, Nomina auf Basis ihres lokalen Kontextes getrennt voneinander zu disambiguieren ohne dabei die Art vorhandener Relationen zu berücksichtigen. Der Ansatz und die sich ergebenden Metriken sind derart entworfen, als dass sich im Anschluss an die WSD durch paarweise Berechnung der prozentualen Überlappung lokal aufgelöster Kontexte automatisch lexikalische Ketten ergeben. Als Grundlage für die Bildung der Kontextmodelle wird *WordNet* als *ground truth* hinsichtlich der Wortbedeutungen verwendet und iterativ mithilfe des semantischen Netzwerkes *ConceptNet* (siehe 2.4.2.) augmentiert.

Die Arbeit ist wie folgt aufgebaut. Zunächst wird in 2. in die relevanten Begriffe, Problemstellungen und verwendete Daten eingeführt, bevor in 3. der Stand der Forschung zusammengefasst und diskutiert wird. In 4. wird der Ansatz beschrieben. Es folgt in 5. eine stichprobenartige Evaluation- und in 6. die Diskussion des Ansatzes. In 7. wird ein Ausblick für weitere Untersuchungen und Nachforschungen gegeben.

2. Einführung

Nachfolgend wird in den Begriff der Textkohäsion, in das Prinzip lexikalischer Ketten und in das Problem der Wortdisambiguierung eingeführt. Anschließend werden *WordNet* und *ConceptNet* vorgestellt und verglichen.

2.1. Textkohäsion

Das Vermögen, einen Text zu verstehen, wird insbesondere durch Textkohärenz und Textkohäsion bestimmt (Halliday und Hasan, 1976). Während Kohärenz den Textzusammenhang auf inhaltlicher, logischer Ebene beschreibt, bezieht sich Kohäsion auf den syntaktischen Zusammenhang. Kohäsion ist damit eine textbildende, semantische Relation. Sie sichert, dass Sätze syntaktisch zusammenhängen oder als zusammenhängend betrachtet werden.

Es gibt verschiedene Kohäsionsmittel, die einen Text als zusammenhängend erkennen lassen. Robuste Indikatoren sind einerseits Wiederholungen (direkte oder partielle Rekurrenz, Synonymie, Proformen und Hyperonomien) und andererseits semantische Relationen, die durch den Kontext mehrerer Wörter bestimmt wird (Halliday und Hasan, 1976). Während die semantische Ähnlichkeit von Wörtern durch Wiederholungen unter Verwendung von *stemming*-Verfahren und hierarchisch organisierten Wortnetzen robust bestimmt werden kann, ist die Registrierung semantischer Bezüge zwischen über den Kontext verbundene Wörter problematisch. Einfache semantische Relationen wie Mengenzugehörigkeiten, Antonomien oder Aggregationen bezeichnet man als systematisch, da sie, wie der Name sagt, relativ einfach abzubilden sind. Nicht-systematische Relationen sind im Umkehrschluss semantische Relationen, die schwer oder nur unzureichend abgebildet werden können, obwohl sie im *common sense* des Menschen existieren, wie z.B. Verbundkonzepte wie „Katzen jagen gerne Mäuse“,

welches in diesem Fall die Wörter „Katze“ und „Maus“ im Kontext der „Jagd“ in Beziehung setzt.

2.2. Lexikalische Ketten

Als lexikalische Ketten bezeichnet man Sequenzen von Wörtern ähnlichen semantischen Bezuges (Morris und Hirst, 1991), die die kohäsive Struktur des zugrundeliegenden Textes widerspiegeln. Einzeln betrachtet, kann man mit ihnen Sätze oder ganze Textpassagen auf übergeordnete, homogene Themenkomplexe abbilden. Lexikalische Ketten können damit, je nach Länge (Anzahl der Kettenelemente), Dichte (räumliche Distanz zwischen Elementen) und Genauigkeit (bezüglich der WSD) indirekte Evidenzen für kohäsive Textfragmente anzeigen. Im Allgemeinen wird ein Text zunächst in Sätze aufgeteilt und tokenisiert, anschließend werden über ein *stemming*-Verfahren die Wortarten und die Lemmata bestimmt. Identifizierte Kandidatenwörter (wie z.B. Nomina) werden dann disambiguiert und paarweise auf Vorkommen von Kohäsionsmitteln überprüft, z.B. mit *WordNet* auf Synonymie und Hyperonomie. Kann ein semantischer Bezug zwischen Kandidaten hergestellt werden, werden die Wörter miteinander verknüpft, sodass sich eine lexikalische Kette ergibt.

Lexikalische Ketten sind eine Art Metadarstellung des Textes und repräsentieren in komprimierter Form dominante Themenkomplexe. Die Art der weiteren Verarbeitung hängt von der umgebenden NLP-Anwendung ab. Beispielsweise kann man durch Tilgung von Textfragmenten, die von Ketten abgedeckt werden, die nicht mit dem Hauptthema korrespondieren, z.B. Textzusammenfassungen in Form von Extrakten ableiten.

2.3. Wortdisambiguierung

Als *word sense disambiguation* (WSD) eines Wortes bezeichnet man im Allgemeinen die automatische Bestimmung der konkreten Wortbedeutung des Wortes anhand seines Kontextes. Diese Fähigkeit ist für das *lexical chaining* wichtig, denn ist die WSD nicht zuverlässig, so ist es unvermeidbar, dass Wörter fälschlicherweise verkettet werden. Beispielsweise kann das Wort „Ball“ sich sowohl auf das Spielgerät, als auch auf die Tanzveranstaltung beziehen. Ein oft verwendeter Disambiguierungsansatz ist, Wortbedeutungen im Kontext so zuzuweisen, dass alle Bedeutungen zueinander passen unter der Annahme, dass für eine Wortform in einem Text jeweils nur eine bestimmte Bedeutung gemeint ist[1]. Beispiel:

> Als Debütantin bezeichnet man eine junge Frau, die auf einem *Ball* „in die Gesellschaft eingeführt“ wird.

Hier wird die „Tanzveranstaltung“ als Bedeutung von „Ball“ angenommen, da die Wörter „Debütantin“ und „Gesellschaft“ Indikatoren hierfür sind. Der Grad der semantischen Ähnlichkeit zweier Wörter wird dann für alle möglichen Bedeutungen über den Kontext abgeschätzt. Anschließend wird jedem Wort eine Bedeutung so zugewiesen, dass die Ähnlichkeit global optimiert wird. Das bedeutet dann für das allgemeine *lexical chaining*, dass das Wort „Ball“ in diesem Satz dementsprechend nicht mit Sportbegriffen wie „Tor“, „Schläger“, „Elfmeter“ oder „Spielfeld“ verkettet werden darf und umgekehrt. Im Allgemeinen ist das Problem der WSD ungelöst; einen Überblick über den Stand der Forschung bietet (Navigli, 2009).

[1] Diese Methodik nennt man auch „one sense per discourse“ (William A. Gale, 1992) und wird in 3.2. aufgegriffen.

People live in houses.
Running is faster than walking.
A person wants to eat when hungry.
Things often found together: lightbulb, contact, glass.
Coffee helps wake you up.
Birds can fly.
The effect of going for a swim is getting wet.
The first thing you do when you wake up is open your eyes.

Tab. 1: Beispiel aus dem OMCS Korpus (Henry Liebermann, 2004, Tab. 1)

2.4. Wissensresourcen

Im Folgenden wird in *WordNet* und *ConceptNet* eingeführt. Gemeinsamkeiten und Unterschiede werden im Anschluss diskutiert.

2.4.1. WordNet

WordNet ist eine etablierte Ressource für NLP-Anwendungen (Fellbaum, 1998). Es ist seit 1985 am *Cognitive Science Laboratory* der *Princeton University* in Entwicklung und wird von Wissensarbeitern händisch erstellt. Es besteht aus einer Datenbank, die konzeptuelle und lexikalische Beziehungen zwischen Wörtern des Englischen abbildet. Es sind Nomen, Verben, Adjektive und Adverben erfasst. *WordNet* wurde nach psycholinguistischen Erkenntnissen entworfen und anhand der Synonymie von Wortformen organisiert: Wörter gleicher Bedeutung sind zu Wort-, bzw. Begriffsfeldern in Form distinktiver Konzeptknoten zusammengefasst, die man *synsets* nennt. Ein *synset* ist die zentrale Repräsentationseinheit innerhalb von *WordNet* und entspricht einem konkreten lexikalischen Sinn. *Synsets* stehen durch ihre semantischen Beziehungen zueinander in Verbindung und sind durch Auszeichnung ihrer gegenseitigen Hyperonomie, bzw. Hyponomie hierarchisch strukturiert. Es existiert keine singuläre Spitze in der Hierarchie, es existieren mehrere abstrakte Wörter, die die Hierarchiespitzen bilden, wie z.B. „thing“ oder „entity“. Durch weitere Relationen wie Antonomie und Meronomie, bzw. Holonomie sind zusätzliche konzeptuelle Relationen erfasst. Die am weitesten reichenden Relationsbäume bestehen zwischen Nomen.

2.4.2. ConceptNet

ConceptNet (Liu und Singh, 2004) wurde vom *MIT Media Laboratory* entwickelt und ist die zur Zeit umfangreichste Datenbank allgemeinen menschlichen Wissens. *ConceptNet* ist ein frei verfügbares, relationales semantisches Netzwerk, welches aus 700.000 englischen Sätzen vom *Open Mind Common Sense* (OMCS) Korpus abgeleitet wurde; vgl. Tab. 1. Anders als in *WordNet*, stellen Knoten in *ConceptNet* Verbundkonzepte in Form von natürlichsprachlicher Fragmente dar, wie z.B. „food“, „buy

food“, „grocery store“ und „at home“. *ConceptNet* wurde mit dem Anspruch entwickelt, allgemeines Weltwissen realitätsnah abzubilden. Hierfür werden 20 semi-formale Relationen verwendet[2], die Konzepte u.A. mit kausalen, räumlichen oder funktionalen Aussagen in Beziehung setzen. *ConceptNet* wurde bereits in NLP-Anwendungen eingesetzt (Liu und Singh, 2004, Kap. 6)(Henry Liebermann, 2004).

2.4.3. Vergleich

WordNet und *ConceptNet* haben folgende Gemeinsamkeiten: beide Resourcen sind relationale, semantische Netzwerke, sind universell anwendbar (nicht domänenspezifisch) und beschreiben Konzepte natürlichsprachlich.

Sie unterscheiden sich in der Form der Wissensbeschaffung, der Systematik und der Behandlung von Ambiguitäten, und zwar wie folgt: Während *WordNet* händisch von Wissensarbeitern erstellt wird, stellt das *ConceptNet* ein automatisch generiertes Derivat des OMCS Korpus dar. Der OMCS Korpus wird durch arbiträre, informale natürlichsprachliche Aussagen zusammengestellt, die jedermann[3] in einem Formular auf der OMCS Website eingeben kann. Allerdings kann man daraus keinen Schluss über die Vertrauenswürdigkeit der Aussagen ziehen (intuitiv erscheint *WordNet* als qualitativ hochwertiger), da die Systematik, bzw. die Zielsetzung beider Resourcen unterschiedlich ist: während der Fokus von *WordNet* eher auf formalen Taxonomien von einzelnen Konzepten liegt, liegt der Fokus von *ConceptNet* eher auf der Erfassung von breitgefächerten Relationen zwischen Verbundkonzepten. Dabei sind in *ConceptNet* ambige Konzepte nicht explizit erfasst wie in *WordNet* über *synsets*, sondern sind nur durch eine Analyse von Konzepten und Relationen auflösbar. Im Gegensatz dazu kann *ConceptNet* als Lösung des „Tennisproblems“[4] (Fellbaum, 1998) verwendet werden, wozu *WordNet* nicht in der Lage ist.

Verglichen mit dem Umfang von *WordNet* wird die Abdeckung von *ConceptNet* in der Literatur (noch) als tendentiell schwach charakterisiert; beispielsweise „weiß“ es, dass eine Ente abgeschossen werden kann, aber nicht, dass dies genauso für einen Fasan gilt. Dieser Umstand ergibt sich dadurch, dass noch keine Person eine Aussage dieser Art in den OMCS Korpus eingegeben hat, oder weil die Aussage nicht durch die automatische Textverarbeitung prozessiert werden konnte. Allerdings kann man diese Information derart interpretieren, dass in solchen Fällen die Wahrscheinlichkeit also höher ist, dass ein Mensch das Abschießen eher mit einer Ente als mit einem Fasan in Beziehung setzt. Deshalb ist das Fehlen von Informationen pauschal nicht per Definition als Mangel anzuerkennen. Aus diesem Grund wird für *commonsense knowledge*-Anwendungen in der Literatur häufig der gemeinsame Einsatz von *WordNet* und *ConceptNet* empfohlen: für eine gegebene Eingabe wird das *synset* in *WordNet* identifiziert und die so ermittelten Synonyme parallel in *ConceptNet* prozessiert; vergleiche (Ming-Hung und Chen, 2006).

3. Stand der Forschung

Lexikalische Ketten wurden durch (Morris, 1988), und später (Morris und Hirst, 1991), eingeführt. Als Wissensresource verwendeten die Autoren Roget's Thesaurus (Roget, 1977), welcher nicht computationell prozessierbar war. Dieser Umstand führte dazu, dass die Autoren ihr Verfahren nicht implementieren konnten und dazu gezwungen waren, die Normalisierung, die Identifizierung von Wortrelationen und das *lexical chaining* per Hand durchzuführen. Aus diesem Grund wird auf die Arbeit von Morris und Hirst lediglich aus historischen Gründen verwiesen[5].

3.1. Relevante Verfahren

Im Folgenden werden die zu diesem Zeitpunkt relevanten Verfahren vorgestellt. Der Schwerpunkt der folgenden Betrachtung liegt in der jeweils angewendeten Methodik. Den Verfahren ist gemein, dass ausschließlich *WordNet* als Wissensresource verwendet wird (vgl. 2.4.1.) und dass ausschließlich Nomen als Kandidatenwörter identifiziert werden[6], entweder über den Abgleich mit *WordNet* (3.1.1., 3.1.4.) oder über den Einsatz eines *part of speech* (POS)-taggers (Brill, 1992) (vgl. 3.1.2., 3.1.3.). Die in den Arbeiten verwendeten Relationen unterscheiden sich, aber alle verwenden Rekurrenzen & Synonyme, identifizierbar über das *synset* eines Wortes. Hyperonomie und Hyponomie-Relationen werden über Pfade in der *WordNet*-Hierarchie abgeleitet. Geschwister-Relationen ergeben sich bei einem gemeinsamen, hyperonymen *synset*, teilweise variiert die Definition; dies wird gesondert angegeben.

3.1.1. Hirst und St.Onge

Das erste computationelle Modell für lexikalische Ketten wurde von (Hirst und St.Onge, 1998) vorgestellt: mit dem in linearer Zeit ausführbaren Algorithmus verarbeitet man in einem Durchlauf den gesamten Text und baut dabei lexikalische Ketten auf, während die Kandidatenwörter zeitgleich disambiguiert werden. Eine Wortinstanz wird genau dann in eine Kette übernommen, wenn eine Relation zu einem Kettenelement existiert; die Menge der Wortbedeutungen beider Elemente wird dann über Schnittmengenbildung reduziert. Die Prüfung auf existente Relationen ist in dieser Rangfolge gestuft: *extra-strong* für Rekurrenz, *strong* für Synonymie, Antonomie & Ähnlichkeit (Kanten

[2]Eine Auswahl ist in 4.2.3. zu finden.

[3]In 2004 waren unter den 100 aktivsten Benutzern des OMCS Projektes u.a. ein Künstler, ein Chemiker, eine Großmutter, ein Rennreiter, ein 12-jähriges Kind und ein Polizist vertreten (Liu und Singh, 2004).

[4]Diese Bezeichnung entstammt der *WordNet*-Gemeinde und bezieht sich darauf, dass im *common sense* ein intuitiver Zusammenhang zwischen den Begriffen „Tennis“, „Ball“, „Schläger“, „Netz“, usw. besteht, aber in *WordNet* nicht durch Relationen wie Hyperonomie/Hyponomie erfasst ist. Zwar existiert ein Pfad zwischen den beteiligten *synsets*, allerdings verläuft dieser durch die abstrakten Wurzel-*synsets* und lässt damit nicht auf einen engen, semantischen Zusammenhang schließen.

[5]Das Verfahren wurde von (Stairmand, 1994) mit einer nun verfügbaren, prozessierbaren Fassung von 1911 ansatzweise umgesetzt. Insgesamt scheiterte das Vorhaben aber, da der Wortindex fehle und das verfügbare Vokabular nicht für die Erfassung zeitgenössischer Texte geeignet sei.

[6](Barzilay und Elhadad, 1997) verwenden zusätzlich Komposita, wie z.B. „swimming pool“.

in *WordNet*), partielle Rekurrenz von Verbundwörtern und *medium-strong* für über Pfade verbundene Wörter (Hyperonomie, Hyponomie, Geschwister, Kontext). Den *medium-strong* Relationen liegen umfangreiche Pfadregeln zugrunden, die der Annahme folgen, dass dadurch semantisch naheliegende Wörter in Relation gesetzt werden können. Sowohl die Autoren als auch andere Forscher beobachten, dass auf diese Weise Wörter (fälschlicherweise) aufgenommen werden, die thematisch nicht mit der Kette korrespondieren. Hirst und St.Onge begründen diese Beobachtung mit der Dichte und der Struktur von *WordNet*; die Arbitrarität des Regelentwurfs in Bezug auf das Tennisproblem (vgl. 2.4.3.) bleibt unerwähnt.

3.1.2. Barzilay und Elhadad

(Barzilay und Elhadad, 1997) weisen auf die vorwärtsgerichtete Propagierung von Wortbedeutungen hin, die durch die von (Hirst und St.Onge, 1998) vorgeschlagene paarweise Schnittmengenbildung resultiert. Dies sei ein konzeptueller Fehler, da dadurch die Akkuranz der WSD maßgeblich von der Reihenfolge der prozessierten Kandidatenwörter abhänge: sich im Verlauf eines Diskurses entwickelnde thematische Schwerpunkte könnten so nicht erfasst werden. Zur Lösung dieses Problems wird vorgeschlagen, 1.) beim Einfügen eines Kettenelements die Wortbedeutungen aller Kettenmitglieder zu erhalten und 2.) die Relationen an konkrete Wortbedeutungen der Kettenelemente zu binden. Zur Auflösung von Disambiguitäten wird eine Bewertungsfunktion eingeführt, mit der ein Interpretations-*score* auf Basis der Homogenität und der Summe von Relationengewichten numerisch berechnet werden kann. Die disambiguierte, konkrete lexikalische Kette ergibt sich dann aus der Bestimmung der Interpretation mit maximalen *score*. Über ein empirisch ermitteltes Relevanzkriterium kann eine Rangfolge der lexikalischen Ketten ermittelt werden.

Da der Bedeutungsraum nicht reduziert wird, ist es zu jedem Zeitpunkt möglich, Wortinstanzen korrekt in eine Kette einzusetzen. Allerdings ergibt sich dadurch, dass alle möglichen Interpretationen berechnet werden, ein exponentieller Speicher- als auch Zeitbedarf, weshalb dieses Verfahren für allgemeine NLP-Anwendungen ungeeignet ist.

3.1.3. Silber und McCoy

Mit dem von (Silber und McCoy, 2002) vorgeschlagenen Verfahren werden für die WSD - ähnlich wie bei (Barzilay und Elhadad, 1997) - alle Kandidaten betrachtet. Der Algorithmus besitzt jedoch eine lineare statt einer exponentiellen Laufzeit. Dies wird durch eine zweistufige Abwärtsstrukturierung möglich: Statt wie bisher Ketten aufgrund von Relationen zu bilden (um daraufhin alle Kombinationen von Wortbedeutungen zu evaluieren), erstellt man zunächst alle möglichen lexikalischen Ketten des Textes. Dies geschieht, indem für alle Kandidatenwörter die jeweils möglichen direkten-, synonymen oder hyperonymen Bedeutungen bestimmt werden und jede Kandidateninstanz mit diesen in Reihenfolge ihres Vorkommens referenziert werden. Silber und McCoy nennen diese Ketten *metachains*, da sie ein mögliches potentielles Thema im Text reflektieren und mögliche potentielle Kandidaten damit in Verbindung bringen; man spricht auch von einem *overriding sense*, den eine *metachain* damit repräsentiert.

Die Disambiguierung der Kandidaten ergibt sich dann aus der impliziten Falsifizierung der *metachain*-Kettenmitglieder: Für jede ambige Wortinstanz wird in Reihenfolge ihres Vorkommens im Text pro assoziierter *metachain* jeweils die Summe der Gewichte der Relationen zu anderen Kettenmitgliedern berechnet. Dazu werden Rekurrenzen, Synonymen, Hyperonymen und exakten[7] Geschwistern ein Gewicht abhängig von der Distanz im Text zugeordnet (im gleichen Satz, im gleichen Abschnitt; drei Sätze entfernt; oder weiter). Der *overriding sense* derjenigen *metachain*, die diese relationale Summe zum betrachteten Kandidaten maximiert, wird der Wortinstanz als konkrete Bedeutung zugewiesen; damit ist die Wortinstanz disambiguiert. Existieren zwei oder mehr *metachains*, die in Frage kommen, wird der niedrigfrequenteste[8] *overriding sense* gewählt. Alle anderen Exemplare der Wortinstanz, die in anderen *metachains* zuvor referenziert wurden, werden dann gelöscht, da sie mit dem *overriding sense* dieser nicht korrespondieren; ist ein zu löschendes Exemplar das letzte verbliebende einer *metachain*, so wird diese ebenfalls gelöscht. Nach Abschluss des Vorgangs sind alle Kandidaten disambiguiert und jeweils genau einer *metachain* zugeordnet. Die verbliebenen metachains sind die nunmehr konkreten, lexikalischen Ketten des Textes, die mit dem Relevanzkriterium von (Barzilay und Elhadad, 1997) ausgewertet werden können.

3.1.4. Galley und McKeown

(Galley und McKeown, 2003) weisen in ihrer Arbeit auf den integrativen Charakter der WSD in allen bisher vorgeschlagenen *lexical chaining* Verfahren hin und stellen die These auf, dass die separate Lösung beider Teilprobleme zu akkurateren Ergebnissen führen würde. Dazu schlagen sie ein in linearer Zeit durchführbares Verfahren vor, mit dem man wie bei Barzilay und Elhadad zunächst das Ziel verfolgt, alle Interpretationen des Textes zu erfassen. Diese werden in einem sogenannten *disambiguation graph* abgebildet, der (gewichtete) Relationen zwischen Wortinstanzen, geordnet nach Wortbedeutungen, erfasst. Dabei wird pro Wortinstanz ein Knoten angelegt und nach seinen möglichen Bedeutungen partitioniert. Eine Relation wird genau dann als Kante zwischen zwei Wortinstanzen hinzugefügt, wenn eine Wortinstanz mit einer seiner Bedeutungen eine Relation zu einer Bedeutung einer anderen Wortinstanz besitzt. Dabei wird der Kante ein Gewicht zugeordnet, welches sich aus dem Relationstyp und der textuellen Distanz zwischen den Wortinstanzen berechnet; die Kante mündet jeweils in den Knotenpartitionen, entsprechend der Bedeutungen, für die sie für die beiden Wortinstanzen definiert ist. Nach Abschluss dieses Vorgangs sind alle Interpretationen des Textes erfasst. Hierfür wird eine lineare Laufzeit angegeben, da das Verfahren von (Silber und McCoy, 2002) in Teilen zugrunde liegt.

[7] Mit exakt ist die gleiche hyperonyme Entfernung zu einem gemeinsamen Vorfahren im Wortnetz gemeint.

[8] Die *synset*-Indizes sind für *WordNet* derart gewählt, als dass sie bei hohem Wert einem niedrigfrequenten Vorkommen im Englischen entsprechen und andersherum.

Galley und McKeown folgen der Annahme „*one sense per discourse*“ (William A. Gale, 1992), also, dass eine hohe Wahrscheinlichkeit besteht, dass ein Wort innerhalb eines Diskurses genau eine Bedeutung entspricht. Man disambiguiert also ein Wort, indem man diejenige Bedeutung ermittelt, für die die Summe der Kantengewichte dieser Bedeutung für alle Instanzen maximal wird. Diese Bestimmung wird für jedes Wort der Kandidatenmenge durchgeführt, und zwar ohne die Struktur des *disambiguation graphs* zu ändern. Auf diese Weise könne, anders als bei Silber und McCoy, keine Propagierung von Wortbedeutungen stattfinden, da der Informationsgehalt pro Kandidat für die WSD nicht durch zuvor getroffene Entscheidungen progressiv reduziert würde. Haben zwei oder mehrere Bedeutungen denselben maximalen *score*, wird die höchstfrequenteste Bedeutung verwendet.

Für die Bestimmung der lexikalischen Ketten werden diejenigen Kanten aus dem *disambiguation graph* entfernt, die durch die WSD keine Gültigkeit mehr besitzen. Dies ist genau dann der Fall, wenn ein Kantenende in eine Knotenpartition mündet, dessen Bedeutung nicht mit der Bedeutung übereinstimmt, die dem Knoten zugewiesen wurde. Der *disambiguation graph* zerfällt dadurch in Teilgraphen. Werden deren Knoten gemäß der Reihenfolge ihres Vorkommens im Text topologisch sortiert, erhält man die lexikalischen Ketten des Textes.

Galley und McKeown evaluieren die Qualität der WSD ihres Verfahrens mithilfe des *semantic concordance corpus* (*SemCor*) (Mihalcea, 1998). Semantische Konkordanzen sind Dokumente, die von Hand mit *WordNet synset* Indizes versehen wurden; die verwendeten 74 Dokumente beinhalten ca. 35.000 Nomen, die zu disambiguieren sind. Galley und McKeown geben eine Akkuranz von 62,09% für ihr Verfahren an. Durch die Evaluation der Verfahren von (Barzilay und Elhadad, 1997) (56,62%) und (Silber und McCoy, 2002) (54,48%) konnten sie ihre These im Rahmen ihrer Evaluation bestätigen.

3.2. Diskussion

Alle Verfahren unterscheiden sich hinsichtlich der WSD. Die Argumentation von (Galley und McKeown, 2003), dass die WSD abgeschlossen sein muss, bevor das eigentliche *lexical chaining* stattfindet, scheint schlüssig; bereits (Barzilay und Elhadad, 1997) kritisierten die progressive Propagierung von Wortbedeutungen bei (Hirst und St.Onge, 1998); dies findet implizit auch bei der Anwendung des Verfahrens von (Silber und McCoy, 2002) statt. Allerdings bietet die Evaluationsmethodik von Galley und McKeown Raum zur Kritik:

One sense per discourse Die Hypothese von (William A. Gale, 1992), dass Wörter in einem Diskurs nur eine Bedeutung haben, korrespondiert nach (Krovetz, 1998) nur für Homonyme (96%), jedoch nicht für Polyseme (66%). Krovetz weist dies u.a. für den *SemCor*-Korpus nach, den auch Galley und McKeown benutzt haben.

Overriding senses Im *SemCor*-Korpus sind nur konkrete Bedeutungen erfasst, sodass *overriding senses* nach Silber und McCoy in der Evaluation von Galley und McKeown nicht registriert werden können. Es bleibt zu prüfen, wie akkurat das Verfahren von Silber und McCoy ist, wenn während der Evaluation *overriding senses* über die Länge ihres Hyperonympfades zur Messung beitragen würden.

Relevanzkriterium Da in Barzilay und Elhadad keine explizite WSD durchgeführt wird und somit theoretisch Wörter mehrfach mit unterschiedlichen Bedeutungen in mehreren Ketten auftreten können, haben Galley und McKeown jeweils als Approximation der WSD die Bedeutung des Wortes in der stärksten Kette verwendet. Da das Relevanzkriterium von Barzilay und Elhadad sich aber über die Anzahl von Relationen und der Homogenität der Elemente der Kette definiert, gibt es keinen unmittelbaren Aufschluss über den Anteil des Wortes an der Kette selbst, sodass unklar ist, ob die Annahme von Galley und McKeown gerechtfertigt ist.

4. Ansatz

Im Folgenden wird ein Ansatz zur Bildung lexikalischer Ketten formuliert, der *common sense* berücksichtigt und für den Einsatz von *WordNet* und *ConceptNet* ausgelegt ist.

4.1. Vorverarbeitung

Der zu verarbeitende Text T wird in Sätze $Q_T = \{q_1, \ldots, q_n\}$ zerlegt. Jeder Satz q wird tokenisiert, die Wortart ermittelt und alle Nicht-Nomina entfernt, sodass sich die Menge der Kandidatenwörter $W_q = \{w_1, \ldots, w_n\}$ eines Satzes q ergibt. Das Lemma eines Kandidatenwortes w wird im Folgenden mit l_w denotiert.

4.2. Kontextmodell

Im Allgemeinen dient das im Folgenden formulierte Kontextmodell der indirekten WSD durch gewichtete Assoziation von *cues* (Signalwörtern) im satzübergreifenden Umfeld einer Wortinstanz w mit den Wortbedeutungen des Kandidatenwortes. Im Folgenden wird das Modell und dessen Metriken formal definiert und anschließend ein Ansatz zur Konstruktion beschrieben.

4.2.1. Formale Darstellung & Metriken

Das Kontextmodell K_l jedes Lemmas l der Kandidatenwörter ist definiert als

$$K_l = (S, C, f_{c,s}) \tag{1}$$

Die Menge $S = \{s_1, \ldots, s_{|S|}\}$ repräsentiert *overriding senses* von l; ein *overriding sense* repräsentiert dabei ein oder mehrere *synsets* in *WordNet*. Bei mehr als einem *synset* generalisiert ein *overriding sense* somit über mehrere konkrete Wortbedeutungen, die in *WordNet* angegeben sind. Die Menge $C = \{c_1, \ldots, c_{|C|}\}$ stellt alle *cues* von l dar, wobei jedes *cue* c selbst als Lemma vorliegt. Die Funktion $f_{c,s}$ ordnet einem *cue* c und einem *overriding sense* s ein *likelihood* zu, dass die Relevanz von c für die Bedeutung s angibt, wenn c im Umfeld von l_w gefunden wird[9].

[9]Ein *likelihood* ist ein statistischer Wert, der eine nichtnormalisierte Wahrscheinlichkeit beschreibt. In diesem Zusammenhang definiert sich das *likelihood* eines *cues* c über die Anzahl von Nennungen in den Wissensresourcen im Kontext von einem Lemma l und einem *overriding sense* s: je häufiger c in diesem Zusammenhang in den Wissensressourcen genannt wird, desto wahrscheinlicher ist es, dass c ein Signalwort für diesen *over-*

$C_s \subseteq C$ ist definiert als relevante Menge[10] von *cues* für s mit $\forall c \in C_s : f_{c,s} > 0$. Die Menge $C_s^* \subseteq C_s$ ist die signifikante Menge[11] von *cues* für s mit

$$c_s^* \in C_s^* : f_{c,s} > \sum_{i=1}^{|C_s|} \frac{f_{c_i,s}}{|C_s|}, \ \forall c_i \in C_s \tag{2}$$

Die Funktion $O_{s,s'}$ berechnet die prozentuale Überlappung der gemeinsamen *cues* $c_j \in J_{s,s'}, J_{s,s'} = C_s \cup C_{s'}$:

$$O_{s,s'} = \frac{\sum_{j=0}^{|J_{s,s'}|} \min(f_{c_j,s}, f_{c_j,s'})}{\sum_{j=0}^{|J_{s,s'}|} \max(f_{c_j,s}, f_{c_j,s'})}, \ \forall c_j \in J_{s,s'} \tag{3}$$

4.2.2. Initialisierung

Ein Kontextmodell K_l wird initialisiert, indem für jedes *synset*, das in *WordNet* für l gespeichert ist, ein *overriding sense* s angelegt wird ($S' \leftarrow S \cup s$). Für jedes *synset* des *overriding sense* s wird anschließend die Glosse und die Wörter des *synsets* tokenisiert & lemmatisiert. Die Nomina werden jeweils als *cue* c' dem Kontextmodell hinzugefügt ($C' \leftarrow C \cup c'$) und das zugehörige *likelihood* von c' & s um 1 erhöht (4), bzw. auf 1 gesetzt, wenn $c' \notin C$ war (5):

$$f'_{c',s} \leftarrow f_{c',s} + 1 \tag{4}$$

$$f'_{c',s} \leftarrow 1 \tag{5}$$

4.2.3. Flache Kontexterweiterung

Das (initialisierte) Kontextmodell K_l wird „flach“ erweitert, indem Konzepte für l in *WordNet* und *ConceptNet* betrachtet werden, die in direkter Relation zu l stehen (eine Kante im semantischen Netz). Zunächst wird für jeden overriding sense s hyponyme und hyperonyme *synsets* in WordNet ermittelt. Diese werden wie in 4.2.2. beschrieben verarbeitet und benutzt, um K_l bezüglich s zu erweitern. Anschließend werden Konzepte in ConceptNet ermittelt, die in unmittelbarer Relation zu l stehen. Es werden die Relationstypen *IsA, ConceptuallyRelatedTo, HasA, AtLocation, HasProperty, MadeOf, UsedFor, PartOf, SymbolOf* jeweils zweimal abgefragt: einmal für konkretisierende-, und einmal für generalisierende Konzepte, die über die genannte Relation mit l in Verbindung stehen. Jedes ermittelte Verbundkonzept $V \equiv Q$ wird tokenisiert & lemmatisiert, Nicht-Nomina verworfen und die übrigen *cue*-Kandidaten c' wie folgt verarbeitet:

- Ist $c' \in C$, wird das *likelihood* um 1 für jeden *overriding sense* s erhöht, für den $c' \in C_s$ ist (4).
- Ist $c' \in C \wedge \exists c'' \notin C$ im selben V ($c', c'' \in V$), dann wird c'' hinzugefügt und das *likelihood* auf 1 für jeden *overriding sense* s gesetzt, für den $c' \in C_s$ ist (5).
- Ist $c' \notin C \wedge c'' \in C$ im selben V ($c', c'' \in V$), dann wird V als unbekanntes *cue*-Konzept für K_l in der neuen Liste Ψ_{K_l} hinzugefügt.

Nachdem K_l erweitert wurde, wird versucht, Bedeutungen zusammenzufassen, die eine hohe Ähnlichkeit aufweisen. Dazu werden einerseits *Related*-Assoziation von *synsets* in *WordNet* und andererseits der Überlappungsgrad von *cue*-Mengen betrachtet. Zwei *overriding senses* werden also vereinigt, wenn eine der beiden folgenden Bedingungen hält: 1.) Existiert in *WordNet* eine *Related*-Assoziation zwischen zwei *synsets*, die in unterschiedlichen *overriding senses* s, s' registriert sind oder 2.) der Überlappungsgrad $O_{s,s'}$ zweier *overriding senses* s, s' größer ist als ein Schwellwert ϑ (z.B. 40%). s, s' werden dann durch den neuen *overriding sense* $s_\cup = s \cup s'$ ersetzt und die *likelihood* Funktion angepasst:

$$\forall c \in C_{s_\cup} : f_{c,s_\cup} \leftarrow f_{c,s} + f_{c,s'} \tag{6}$$

4.2.4. Tiefe Kontexterweiterung

Das (flach erweiterte) Kontextmodell K_l wird „tief“ erweitert, indem zunächst für jedes signifikante *cue* $c^* \in C^* = \cup_{\forall s \in S} C_s^*$ ein flaches Kontextmodell K_{c^*} ermittelt wird und anschließend K_l für jeden *overriding sense* s wie folgt beschrieben geeignet erweitert wird: Für jedes signifikante *cue* $c_s^* \in C_s$ wird derjenige *overriding sense* $s^* \in S$ aus K_{c^*} ermittelt, für den die Überlappung O_{s,s^*} am größten ist. Ist ein *cue* $c' \in C_{s^*}$ nicht in C_s enthalten, wird es K_l hinzugefügt und das *likelihood* mit einem diskontierten Faktor von 0,5 gesetzt (8); ist aber ein *cue* $c' \in C_{s^*} \in C_s$ enthalten, erhöhe das *likelihood* um einen diskontierten Faktor von 0,5 (7):

$$f'_{c',s} \leftarrow f_{c',s} + 0,5 \tag{7}$$

$$f'_{c',s} \leftarrow 0,5 \tag{8}$$

Anschließend wird geprüft, ob *cues* in C_{s^*} existieren, die mit Verbundkonzepten V in α_{K_l} korrespondieren. Ist dies der Fall, werden die zuvor unbekannten *cues* nach den Regeln in 4.2.3. K_l hinzugefügt und aus α_{K_l} gestrichen. Abschließend werden *overriding senses* vereinigt, falls möglich (vgl. 4.2.3.).

4.3. Word sense disambiguation

Die Disambiguierung eines Wortes w im Satz $q_i \in Q_T$ des Textes T geschieht unmittelbar durch die Analyse der lemmatisierten Nomina in der durch die Satzdistanz λ charakterisierte Umgebung $\xi_w = (q_{i-\lambda}, \ldots, q_i, \ldots, q_{i+\lambda})$. Ziel ist es, einen *overriding sense* $s \in S$ von K_l für das Lemma l_w des Wortes w zuzuweisen. Dies wird als *maximum likelihood* Problem behandelt, indem zunächst die *likelihoods* von im Text auftauchenden *cues* je nach *overriding sense* s und abhängig von der Satzdistanz und dem Diskontierungsfaktor δ aufsummiert werden (L_s):

$$L_s^w = \sum_{i=-\lambda}^{\lambda} \sum_{j=0}^{|q_i|} \frac{f_{l_{w_j},s}}{\delta^{|i|}} \tag{9}$$

L_s^w entspricht dem *likelihood* aller *cues* in der Umgebung ξ_w von w für die Bedeutung $s \in S$ des Kontextmodells K_{l_w}

riding sense darstellt. Wird also c im Umfeld von l gefunden und ist ein hohes likelihood für einen overriding sense s zugewiesen, ist also im Umkehrschluss die Wahrscheinlichkeit hoch, dass diese Bedeutung s für die das prozessierte Wort wahrscheinlich ist.

[10] Die *cues* wurden also mindestens einmal in den Wissensressourcen für den *overriding sense* s genannt.

[11] Ein *cue* c ist genau dann für s signifikant, wenn $f_{c,s}$ über dem mittleren *likelihood* der Relevanzmenge C_s liegt.

des Lemmas l der Wortinstanz w. $|q_i|$ entspricht der Anzahl der nominalen Kandidatenwörter im Satz $q_i \in Q_T$ im Text T. δ ist der konstante Basisfaktor der Satzdiskontierung[12] zwischen dem Lemma des betrachteten Kandidatenwortes und dem zu disambiguierenden Wort w.

Diejenige Bedeutung mit größem akkumulierten *likelihood* L_s^* entspricht dann derjenigen (generalisierten) Wortbedeutung s_w^* mit höchster Evidenz und wird der Wortinstanz w dann zugewiesen:

$$s_w^* = \arg\max_s L_s^w \tag{10}$$

4.4. Lexical chaining

Für jedes mit K_{l_w} disambiguierte Kandidatenwort w eines Textes T kann durch die Berechnung der nach Satzdistanz diskontierten paarweisen Überlappung zu einer anderen Wortinstanz w' das relationale *likelihood* $\chi_{w,w'}$ berechnet werden:

$$\chi_{w,w'} = \frac{O_{s_w^*, s_{w'}^*}}{\gamma^{|i-j|}} \tag{11}$$

wobei w, w' die zwei betrachteten Wortinstanzen und $s_w^*, s_{w'}^*$ die nach der WSD zugewiesenen *overriding senses* sind. Die Satznummern i, j der Sätze q_i, q_j im Text T werden zur Satzdiskontierung über den konstanten Faktor γ verwendet, damit die *likelihoods* weit entfernter Kandidaten weniger in das relationale *likelihood* χ eingehen ($\gamma << \delta$), als bei nahe beieinander liegenden Kandidaten.

Das relationale likelihood $\chi_{w,w'}$ drückt also auf numerische Weise aus, wie wahrscheinlich es ist, dass die beiden (nach s und s' disambiguierten) Wörter w, w' eine direkte Relation besitzen; die Art der Relation ist hier nicht ersichtlich. Verbindet man eine Wortinstanz w mit derjenigen Wortinstanz w^*, die χ_{w,w^*} maximiert, ergeben sich starke, lexikalische Ketten.

5. Evaluation

Der in 4.3. skizzierte Ansatz zur *word sense disambiguation* wird im Folgenden anhand ausgewählter Stichproben evaluiert. Dies geschieht mithilfe der *brown1*- und *brown2*-Konkordanzen des *SemCor* Korpus; alle folgenden Aussagen beziehen sich auf beide Konkordanzen, Nominalformen und *synsets* in *WordNet*, die für Nominalformen angelegt sind.

Der Korpus beinhaltet 11536 verschiedene Nominalformen; das Wort „person“ kommt mit 6696 Vorkommen am meisten vor (Rang 1), das Wort „head“ ist diejenige Nominalform mit den meisten nominalen *synsets* (33) und kommt 179 mal vor (Rang 22). Es existieren 83 Nominalformen ohne Korrespondenz mit *WordNet* und 5880 Nominalformen, die nur ein *synset* in *WordNet* zugewiesen haben; es ergeben sich daher 5573 ambige Nominalformen

[12] Ist $\delta = 1$, wird nicht diskontiert, alle *cues* gehen gleichermaßen mit ihrem *likelihood* in L ein. Ist $\delta > 1$, werden die *likelihoods* exponential ausgehend vom Ursprungssatz diskontiert; je größer δ also ist, desto weniger Evidenz bieten *cues* am Rande von ξ. Wird $0 < \delta < 1$ gewählt, werden *cues* zum Rand von ξ stärker gewichtet, was nicht empfehlenswert ist. L ist für $\delta = 0$ nicht definiert.

Rang_{ω_l}	l	N_l	Rang_{N_l}	S_l
1	way	298	8	12
2	man	648	4	11
3	time	511	5	10
4	day	331	7	10
5	location	993	3	4
6	person	6696	1	3
7	group	1329	2	3
8	thing	271	10	12
9	form	232	12	16
10	head	179	22	33
256	bank	37	435	10
285	dog	39	398	7
1253	orange	8	1937	5

Tab. 2: Im oberen Teil sind die ersten zehn absteigend nach $\omega_l = \frac{1}{S_l}^{N_l}$ sortierten lemmatisierten Nominalformen (l) in der *brown1*- und *brown2*-Konkordanz des *SemCor* Korpus aufgeführt; darunter die zusätzlichen evaluierten Lemmata. Die Vorkommen (N_l) und die Anzahl zugewiesener *synsets* (S_l) beziehen sich auf Nominalformen. Der Rang_{N_l} gibt den Rang in der absteigenden Sortierung nach N_l an. Die ω_l-Werte sind nicht abgedruckt, da sie hier nicht darstellbar sind ($\omega_l \to 0$).

(48,31%). Ist N_l die Anzahl der Vorkommen einer Nominalform l in den Konkordanzen und S_l die Anzahl der zugewiesenen nominalen *synsets* in *WordNet*, ergibt sich daher die Wahrscheinlichkeit

$$\omega_l = \left(\frac{1}{S_l}\right)^{N_l} = {S_l}^{-N_l} \tag{12}$$

dass eine Nominalform über die gesamte Konkordanz zufällig korrekt disambiguiert wird; ω_l ist daher eine Art Indikator für die Schwierigkeit der Ermittlung der Akkuranz einer WSD für ein bestimmtes Lemma l in einem Korpus. Es werden im Folgenden einerseits die Nomen „way“, „man“ und „time“ als Repräsentanten schwierig zu evaluierender Lemmata gegen die *brown1*- und *brown2*-Konkordanz evaluiert; anderserseits werden Nomen mit wenige *synsets* „bank“, „dog“ und „orange“; die Ausgangsdatenlage ist in Tab. 2 dargestellt.

Da per Definition ein Kontextmodell K_l mehrere WordNet-synsets zu overriding senses zusammenfassen kann, werden zwei Akkuranz-Maße berechnet. Wird ein in der Konkordanz eingetragenes *synset* mit einem *overriding sense* getroffen, in dem das *synset* beinhaltet ist, dann geht der Treffer mit $1/S_s$ in die (präzise) *accuracy* A_p ein, wobei S_s die Anzahl der *synsets* im *overriding sense* s bezeichnet. Diese Art der accuracy respektiert die Tatsache, dass zwar der overriding sense das zu treffende synset beinhaltet, da aber über mehrere synsets generalisiert wird, dieses nur mit einem Bruchteil in die Akkuranz eingeht. Andererseits ergibt sich, wenn man die Tatsache betrachtet, dass ein overriding sense homogene synsets zusammenfasst, eine nach distinktiven Wortbedeutungen durchgeführte WSD. Aus diesem Grund wird zusätzlich die (generalisierende) *overriding accuracy* A_o berechnet, in die, sobald das in der Konkordanz angegebene synset mit einem synset im gewählten overriding sense korrespondiert, ein „voller“ Treffer eingeht.

l	$\|S\|$	baseline $\|S\|^{-1}$	A_p (flach)	A_p (tief)
way	12	8%	19%	13%
man	11	9%	6%	14%
time	10	10%	7%	9%
bank	10	10%	29%	21%
dog	7	14%	20%	50%
orange	5	20%	31%	37%

Tab. 3: Ergebnisse für die *accuracy* A_p für flache- und tiefe Kontextmodelle. Die *baseline* berechnet sich über die Anzahl von *synsets* für das Lemma l.

l	$\|S\|$	baseline $\|S\|^{-1}$	A_o (flach)
way	12	8%	19%
man	9	11%	6%
time	10	10%	7%
bank	8	13%	59%
dog	7	14%	20%
orange	3	33%	62%

Tab. 4: Ergebnisse für die *overriding accuracy* A_o für flache Kontextmodelle. Die *baseline* berechnet sich über die Anzahl der nach der flachen Kontexterweiterung übrigen *overriding senses*.

Die oben genannten Lemmata werden für die Akkuranzen A_p und A_o einmal für die flache Kontexterweiterung und einmal für die tiefe Kontexterweiterung evaluiert. Es werden folgende Parameter verwendet: $\vartheta = 40\%$, $\delta = 1,1$ und $\lambda = 3$. Es werden für das jeweils verwendete Kontextmodell die Anzahl von *overriding senses* beobachtet und die *baseline* angegeben, mit der pro Lemma die Wahrscheinlichkeit besteht, das in der Konkordanz angegebene synset zufällig zu treffen.

In Tab. 3 sind die Ergebnisse für die *accuracy* A_p zu finden. Die Ergebnisse für die *overriding accuracy* A_o sind aufgeteilt nach Verwendung des flachen Kontextes (Tab. 4) und des tiefen Kontextes (Tab. 5).

Es ist zu beobachten, dass tiefe Kontextmodelle mit den verwendeten Parametern eine *overriding accuracy* A_o haben, die die jeweils berechnete baseline übertreffen - in manchen Fällen („man“, „dog“) wird nach diesem Maß annähernd perfekt disambiguiert. Für die *accuracy* A_p und A_o mit flachen Kontextmodellen sind schwankende Werte zu beobachten. Teilweise liegen die Messergebnisse unter der *baseline*, was darauf schließen lässt, dass in diesen Fällen das Kontextmodell explizit andere Wortbedeutungen wählt, als die Konkordanz dies vorgibt. In Fällen mit ursprünglichen vielen synsets lässt sich beobachten, dass wenn eine starke Reduzierung stattfindet („man“, „orange“) dass relativ hohe Akkuranzwerte erzielt werden können.

l	$\|S\|$	baseline $\|S\|^{-1}$	A_o (tief)
way	12	8%	13%
man	6	17%	88%
time	9	11%	18%
bank	8	13%	43%
dog	6	17%	100%
orange	3	33%	75%

Tab. 5: Ergebnisse für die *overriding accuracy* A_o für tiefe Kontextmodelle. Die *baseline* berechnet sich über die Anzahl der nach der tiefen Kontexterweiterung übrigen *overriding senses*.

6. Diskussion

Der Ansatz basiert auf der Annahme, dass semantische Informationen, die dem Modell hinzugefügt werden, der Datenlage entspricht, die sich in echten Texten wiederfinden und sich die Modellierung der *likelihood*-Funktion an der Frequenz der Vorkommen in den Wissensressourcen orientiert.

Das Problem jedoch ist, dass die iterative Erweiterung des Kontextmodells eine Art Diffusion von Kontextinformationen bewirkt, indem viele neue *cues* und *likelihoods* zum Modell hinzukommen, die zwar mit einem signifikanten *cue* eines Wortes in Verbindung stehen, aber nicht direkt mit dem Wort selbst im *common sense* in Verbindung gebracht wird. Ein weiteres Problem ist, dass *ConceptNet* an sich keine Wortbedeutungen speichert, sodass die „naive“ Zuordnung, wie sie hier stattfindet, oftmals *cues* übrig lässt, die nicht zugeordnet werden können. In diesem Ansatz werden *cues* zudem ohne weitere Analyse aufgenommen, was bei einer auseinandergehenden Abdeckung von *ConceptNet* und *WordNet* angebracht wäre: beispielsweise ist in *ConceptNet* das Wort „force“ mit dem Verbundkonzept (Luke) „sky walker“ verbunden; dies führt hier zu einer möglichen Aufnahmen von „sky“, obwohl das Konzept hier auf eine Gegebenheit in einem Film bezieht, die im *common sense* vorhanden ist, aber nicht von *WordNet* aufgegriffen wird.

7. Fazit und Ausblick

Generell erscheint es sinnvoll, Relationen zwischen Kandidatenwörtern über *likelihoods* zu ermitteln, die sich aus Auftreten in Wissensressourcen ergeben, ohne dabei den Typ der Relation zu betrachten. Allerdings zeigt sich auch, dass eine naive Aufnahme von Informationen problematisch ist; es ist zu klären, welchen Einfluss die verwendeten Relationstypen auf die Akkuranz der WSD haben. Ist bekannt, welche Relationstypen nützlich sind und welche nicht, kann man gezielt likelihoods diskontieren, um stabilere Ergebnisse zu erzielen. Es ist zu klären, ob der Ansatz für die Bildung lexikalischer Ketten geeignet ist.

8. Referenzen

Barzilay, Regina und Michael Elhadad, 1997. Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, Band 17.

Brill, Eric, 1992. A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing*. Association for Computational Linguistics.

Fellbaum, Christiane, 1998. *WordNet: An electronic lexical database*. MIT Press.

Galley, Michel und Kathleen McKeown, 2003. Improving word sense disambiguation in lexical chaining. In *International Joint Conference on Artificial Intelligence*, Band 18.

Halliday, M. A. K. und Ruqaiya Hasan, 1976. *Cohesion in English*. Longman Group Ltd.

Henry Liebermann, Push Singh Barbara Barry, Hugo Liu, 2004. Beating common sense into interactive applications. *AI Magazine*, 25(4):63–76.

Hirst, Graeme und David St.Onge, 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum (Hrsg.), *WordNet: An electronic lexical database*, Kapitel 13. MIT Press.

Krovetz, Robert, 1998. More than one sense per discourse. Technical report, NEC Research Institute, Princeton.

Liu, H. und P. Singh, 2004. Conceptnet a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Mihalcea, Rada, 1998. Semcor semantically tagged corpus.

Ming-Hung, Ming-Feng Tsai und Hsin-Hsi Chen, 2006. Query expansion with conceptnet and wordnet: An intrinsic comparison. *Information Retrieval Technology*:1–13.

Morris, Jane, 1988. *Lexical cohesion, the thesaurus, and the structure of text*. Masterarbeit, University of Toronto.

Morris, Jane und Graeme Hirst, 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of the text. *Computational Linguistics*, 17:21–43.

Navigli, Roberto, 2009. Word sense disambiguation - a survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

Roget, Peter, 1977. *Roget's International Thesaurus*, Band 1. Harper and Row, 4. Auflage.

Silber, H. Grogory und Kathleen F. McCoy, 2002. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics*, 28(4):487–496.

Stairmand, Mark, 1994. Lexical chains, wordnet and information retrieval. Centre for Computational Linguistics, UMIST, Manchester.

William A. Gale, David Yarowsky, Kenneth W. Church, 1992. One sense per discourse. In *Proceedings of the 4th ARPA Workshop on Speech and Natural Language Processing*. AT&T Bell Laboratories, NY: Harriman.