

# Berechnung semantischer Ähnlichkeit: Ein Vergleich zwischen WordNet und FrameNet

Sven Arweiler  
Matrikelnummer 2018802  
svar@coli.uni-saarland.de

Universität des Saarlandes  
Fachrichtung 4.7  
Allgemeine Linguistik

30. Juni 2008

# Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>2</b>
<b>2</b>	<b>WordNet und FrameNet: Ein struktureller Vergleich</b>	<b>4</b>
2.1	Grundbegriffe . . . . .	4
2.2	WordNet . . . . .	5
2.2.1	Relationen in WordNet . . . . .	5
2.3	FrameNet . . . . .	7
2.3.1	Relationen in FrameNet . . . . .	8
2.4	Struktureller Vergleich . . . . .	11
<b>3</b>	<b>Methoden zur Berechnung semantischer Ähnlichkeit</b>	<b>14</b>
3.1	Strukturbasierte Methoden . . . . .	14
3.1.1	Leacock/Chodorow (1998) . . . . .	14
3.1.2	Wu/Palmer (1994) . . . . .	15
3.1.3	Unterschiede zwischen Leacock/Chodorow und Wu/Palmer	16
3.2	Informationsbasierte Methoden . . . . .	16
3.2.1	Resnik (1995) . . . . .	17
3.2.2	Lin (1998) . . . . .	18
<b>4</b>	<b>Anwendung der Methoden auf FrameNet</b>	<b>19</b>
4.1	Anwendung von Leacock/Chodorow auf FrameNet . . . . .	19
4.2	Anwendung von Wu/Palmer auf FrameNet . . . . .	21
<b>5</b>	<b>Evaluation</b>	<b>24</b>
5.1	Wortpaare . . . . .	24
5.1.1	Leacock und Chodorow . . . . .	25
5.1.2	Wu und Palmer . . . . .	26
5.1.3	Diskussion . . . . .	26
5.2	RTE . . . . .	27
5.2.1	Anwendung der Methoden auf RTE . . . . .	27
5.2.2	Diskussion . . . . .	29
<b>6</b>	<b>Fazit</b>	<b>30</b>
6.1	Zusammenfassung . . . . .	30
6.2	Ausblick . . . . .	30
<b>7</b>	<b>Bibliographie</b>	<b>32</b>

# 1 Einleitung

Mittels struktureller Berechnungen kann der Ähnlichkeit der Bedeutung zweier Begriffe durch ein Maß Ausdruck verliehen werden. Dieses Maß wird *semantische Ähnlichkeit* genannt. Semantische Ähnlichkeit findet in verschiedenen Bereichen Anwendung, wie beispielsweise beim Finden von lexikalischen Ketten, in Wortverwechslungsszenarien, bei der Textzusammenfassung oder beim Erkennen von figurativer Sprache. Für die Berechnung der semantischen Ähnlichkeit existieren verschiedene Methoden, die größtenteils auf strukturierte, semantische Lexika wie WordNet<sup>1</sup> zurückgreifen und dem Maß der semantischen Ähnlichkeit durch einen numerischen Wert Ausdruck verleihen. In WordNet bilden Wörter mit synonymem Bedeutung eine Einheit: das *Synset*. Die Synsets sind durch verschiedene Relationen, wie z.B. Hyponym- und Meronymrelation, miteinander verbunden. Außerdem ist ein großer Teil der Synsets durch Glossen und Beispielsätze beschrieben.

Semantische Ähnlichkeit kann als eine speziellere Variante der semantischen Verwandtschaft (*semantic relatedness*) angesehen werden (Budanitsky & Hirst 2006). Bei der Berechnung semantischer Ähnlichkeit wird tatsächlich der inhaltlichen Ähnlichkeit zweier Begriffe ein Wert zugewiesen. So sind beispielsweise die Begriffe *Arzt* und *Krankenschwester* semantisch ähnlich, die Begriffe *Arzt* und *Krankenhaus* aber nicht. Nichtsdestotrotz stehen die Begriffe *Arzt* und *Krankenhaus* in einem semantisch verwandtschaftlichen Verhältnis.

Im Folgenden wird der Begriff der semantischen Ähnlichkeit verwendet werden, da dies den Konzepten der verwendeten Verfahren entspricht.

Semantische Ähnlichkeit approximiert über strukturelle Berechnung (z.B. über Pfadlänge oder Tiefeninformationen in einer Baumstruktur). Ein potentieller Nachteil von WordNet ist dabei, dass Wörter verschiedener Wortarten nicht miteinander verknüpft sind, das heißt es kann keine semantische Ähnlichkeit über Wortartgrenzen hinaus gemessen werden. An diesem Punkt hat das semantische Netz von FrameNet<sup>2</sup> einen entscheidenden Vorteil: Die Frames (idealisierte Situationsbeschreibungen) sind wortartunabhängig organisiert, und somit ist eine Messung der semantischen Ähnlichkeit über Wortartgrenzen hinaus möglich. Zudem sind die Frames durch semantisch aussagekräftige Relationen (z.B. Inheritance, Using) miteinander verbunden. Zum Thema der Ermittlung semantischer Ähnlichkeit auf Basis von FrameNet wurden bisher allerdings keine Arbeiten veröffentlicht.

In dieser Arbeit soll zunächst ein struktureller Vergleich zwischen WordNet und FrameNet vorgenommen werden, und schließlich soll geprüft werden, ob

---

<sup>1</sup><http://wordnet.princeton.edu/>

<sup>2</sup><http://framenet.icsi.berkeley.edu/>

Methoden zur Ermittlung semantischer Ähnlichkeit zwischen zwei Konzepten, die in WordNet Anwendung finden, auch auf FrameNet übertragbar sind. Dabei sollen im Speziellen die Methoden von Leacock und Chodorow (1998) und von Wu und Palmer (1994) betrachtet werden. Leacock und Chodorow wählen den Ansatz einer normalisierten Pfadlänge, das heißt, die Länge des Pfades zwischen zwei Konzepten wird im Verhältnis zur maximalen Tiefe, die in WordNet erreicht werden kann, berechnet. Wu und Palmer hingegen suchen eine *lowest superordinate* von zwei Konzepten, also einen Knoten, der die Eigenschaften beider Konzepte subsumiert.

Weiterhin soll durch das Übertragen dieser Methoden auf FrameNet geprüft werden, ob die bisher nur in WordNet angewandten Methoden in FrameNet ähnliche oder womöglich sogar bessere Ergebnisse liefern können.

Eine Voraussetzung für die Anwendung der Methoden von Wu/Palmer, sowie in geringerem Maße auch bei der Methode von Leacock/Chodorow, ist eine Baumstruktur des zugrunde liegenden Lexikons. Da die Struktur von FrameNet nicht baumartig ist, muss vor der Übertragung des Verfahrens auf FrameNet geprüft werden, ob es durch geeignete Maßnahmen möglich ist, eine baumnahe Struktur zu bilden bzw. ein baumartiges "Rückgrat" zu identifizieren. Desweiteren soll untersucht werden, ob es sinnvoll ist, eine Gewichtung für die Relationen in FrameNet zu vergeben, da nicht alle Relationen die gleiche Relevanz für die Ermittlung der semantischen Ähnlichkeit haben. Bei der Anwendung der Methoden von Leacock/Chodorow und Wu/Palmer spielen die Hypernym- und Hyponym-WordNet-Relationen (*is-a*) eine grundlegende Rolle, das heißt, es muss eine Entsprechung dafür in FrameNet gefunden werden.

Eine Evaluation der FrameNet-basierten Berechnungen wird durch den Vergleich mit von Menschen auf semantische Ähnlichkeit bewerteten Wortpaaren (Rubenstein und Goodenough 1965) und den Werten, die durch die Methoden von Leacock/Chodorow und Wu/Palmer in WordNet ermittelt werden, stattfinden, sowie in einem größeren Rahmen in einem RTE-Szenario (*recognising textual entailment*), in dem entschieden werden soll, ob eine Hypothese aus einem Text geschlussfolgert werden kann.

Zur technischen Umsetzung steht eine Ruby-Implementation des FrameNet-Graphen und eine Schnittstelle zu WordNet zur Verfügung, die von mir entwickelt wurden. Die Ähnlichkeitsdaten für Wortpaare in WordNet können mittels des `WordNet::Similarity-Paketes`<sup>3</sup> für Perl (Pedersen 2004) ermittelt werden, in dem auch die Vergleichsmethoden von Leacock/Chodorow und Wu/Palmer implementiert sind. Die FrameNet-Umsetzung dieser Methoden wurde im Rahmen dieser Arbeit mit einer Ruby-Implementation realisiert.

---

<sup>3</sup><http://wn-similarity.sourceforge.net/>

## 2 WordNet und FrameNet: Ein struktureller Vergleich

Im Folgenden werden die beiden lexikalisch-semantischen Netze WordNet und FrameNet vorgestellt und ein struktureller Vergleich zwischen beiden Ressourcen im Hinblick auf die Berechnung semantischer Ähnlichkeit angestellt. Beide Netze weisen Merkmale auf, die sie vom jeweils anderen deutlich abgrenzen, jedoch sind Ähnlichkeiten vorhanden, die das in dieser Arbeit zu untersuchende Übertragen von Methoden sinnvoll und interessant erscheinen lassen.

### 2.1 Grundbegriffe

Hier werden einige Grundbegriffe kurz definiert, die in dieser Arbeit häufig Anwendung finden werden.

**Graph** Ein Graph  $G$  besteht aus einer Menge von Knoten  $V$  und einer Menge von Kanten  $E$ , die die Knoten von  $G$  verbinden.

**Gerichteter Graph** Ein gerichteter Graph ist ein Graph  $G$ , mit einer Menge von Knoten  $V$  in dem die Richtung der Kanten  $E$  vorgegeben ist. Das heißt eine Kante  $(v, v')$  ist nicht gleich der Kante  $(v', v)$ , wobei gilt  $v, v' \in V$ .

**Eingangsgrad** Der Eingangsgrad eines Knotens  $v \in V$  in einem gerichteten Graph  $G$  ist die Anzahl der Vorgänger von  $v$  in  $G$ , das heißt die Anzahl der Knoten  $v' \in V$ , für die eine Kante  $(v', v) \in E$  existiert.

**Ausgangsgrad** Der Ausgangsgrad eines Knotens  $v \in V$  in einem gerichteten Graph  $G$  ist die Anzahl der Nachfolger von  $v$  in  $G$ , das heißt die Anzahl der Knoten  $v' \in V$ , für die eine Kante  $(v, v') \in E$  existiert.

**Baum** Ein Baum ist ein azyklischer, zusammenhängender Graph, in dem kein Knoten einen Eingangsgrad größer als 1 hat.

**Wurzel** Die Wurzel eines gerichteten Baumes ist der Knoten, von dem alle anderen Knoten erreichbar sind, der selbst aber von keinem Knoten erreichbar ist. Der Wurzel eines gerichteten Baumes hat einen Eingangsgrad von 0.

{ <i>entity</i> }
{ <i>abstraction</i> }
{ <i>psychological feature</i> }
{ <i>natural phenomenon</i> }
{ <i>activity</i> }
{ <i>event</i> }
{ <i>group</i> }
{ <i>location</i> }
{ <i>possession</i> }
{ <i>shape</i> }
{ <i>state</i> }

Tabelle 2.1: Die 11 *unique beginners* der Nomentaxonomie in WordNet

## 2.2 WordNet

WordNet (Fellbaum 1998) ist ein strukturiertes lexikalisch-semantisches Netz für die englische Sprache, das 1985 unter der Leitung des Psychologieprofessors George A. Miller an der Princeton University ins Leben gerufen wurde. WordNet kann sowohl online<sup>1</sup>, als auch in einer lokalen Installation genutzt werden. In WordNet sind Lemma mit der gleichen semantischen Bedeutung (Synonyme) in sogenannten *Synsets* erfasst, die größtenteils durch eine kurze Definition (*gloss*) und einen Beispielsatz beschrieben werden. Das Synset ist der Grundbaustein von WordNet. In der momentan<sup>2</sup> aktuellen Version 3.0, die auch in dieser Arbeit zum Einsatz kommt, enthält WordNet 146312 Nomen, die in 117798 Synsets organisiert sind, 25047 Verben in 11529 Synsets, 30002 Adjektive in 21479 Synsets und 5580 Adverbien in 4481 Synsets<sup>3</sup>.

Die verschiedenen Wortarten liegen in separaten Taxonomien vor. Die Nomentaxonomie ist wiederum in 11 Taxonomien aufgeteilt (Budanitsky & Hirst 2006), deren Wurzelknoten als die *unique beginners* bezeichnet werden. Jeder *unique beginner* repräsentiert ein elementares, allgemein gehaltenes Nomenkonzept, unter dem sich speziellere Konzepte einordnen lassen. Tabelle 2.1 zeigt die 11 *unique beginner*-Synsets (Hayes 1999).

### 2.2.1 Relationen in WordNet

Die Synsets innerhalb der Wortarttaxonomien sind untereinander durch verschiedene semantisch aussagekräftige Relationen verbunden. Relationen zwischen Synsets in WordNet sind Hyperonomie, Hyponomie, Meronymie und Holonymie. Neben Relationen zwischen Synsets existieren lexikalische Relationen, die Beziehungen zwischen Lemmata darstellen. Lexikalische Relationen in WordNet sind Synonymie und Antonymie.

**Synonymrelation** Durch die Synonymrelation werden die Synsets definiert, die “Bausteine” von WordNet, die durch die semantischen Relationen verbunden sind. Wörter gelten als synonym, wenn ein Wort in einem Satz

<sup>1</sup>Web-Oberfläche für WordNet: <http://wordnet.princeton.edu/perl/webwn>

<sup>2</sup>Stand: 2. Juni 2008

<sup>3</sup>WordNet-Statistik: <http://wordnet.princeton.edu/man/wnstats.7WN>

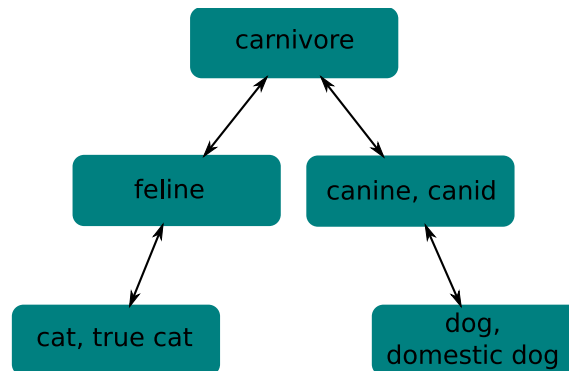


Abbildung 2.1: Hypernym- und Hyponymrelationen in WordNet

innerhalb eines Kontexts durch ein anderes Wort ersetzt werden kann, ohne dass sich der Wahrheitswert des Satzes ändert.

**Antonymie** Die Antonymierelation, ebenfalls eine lexikalische Relation, zeichnet die Gegensätzlichkeit von Wörtern aus. Gegensätzlichkeit bedeutet nicht zwingend, dass *nicht-X* das Antonym von *X* ist. So sind *groß* und *klein* Antonyme, allerdings ist *nicht-groß* nicht gleichbedeutend mit *klein*, denn nicht alles was nicht groß ist, ist klein. Ein Fall, in dem *nicht-X* das Antonym von *X* ist, sind die Wörter *wahr* und sein Antonym *falsch*: Im Allgemeinen kann man davon ausgehen, dass etwas, das nicht wahr ist, falsch ist. Man kann also beobachten, dass *nicht-X* das Antonym vom *X* ist, wenn es sich nicht um Wörter mit relativer, sondern mit aboluter Bedeutung handelt.

**Hyponymrelation** Die Hyponymrelation (auch *is-a*-Relation) ist formal folgendermaßen definiert (Miller 1993):

*X ist ein Hyponym von Y, wenn X eine Art von Y ist.*

Die Hypernymrelation zeigt in die entgegengesetzte Richtung:

*X ist ein Hypernym von Y, wenn Y eine Art von X ist.*

Abbildung 2.1 zeigt die Relationen anhand eines Beispiels mit den Synsets  $\{cat, true\ cat\}$  und  $\{dog, domestic\ dog\}$ . Die Abwärtsrichtung entspricht der Hyponymrelation, vom Allgemeinen zu Speziellen. Die Aufwärtsrichtung stellt die Hypernymrelation dar, vom Speziellen zum Allgemeinen. Da es in fast allen Fällen nur ein Hypernym zu einem Synset gibt, stellt die Hyponymie eine Baum- bzw. Waldstruktur dar. Außerdem stellt die Hyponymrelation eine transitive Vererbungshierarchie dar (Miller 1993), d.h. jedes Synset erbt die semantischen Eigenschaften seines Hypernyms. So müssen beispielsweise auf alle Hyponyme des Synsets  $\{carnivore\}$  die semantischen Eigenschaften von *carnivore* (Fleischfresser), die durch seine Glosse definiert werden, zutreffen.

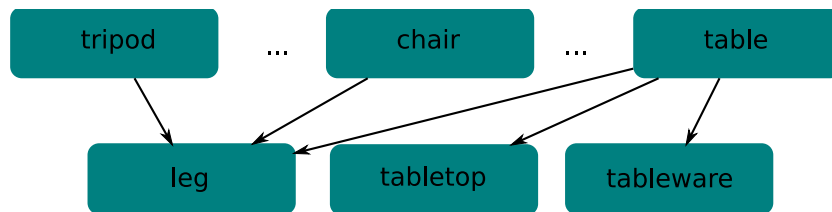


Abbildung 2.2: Meronymrelationen in WordNet

**Meronymrelation** Die Meronymrelation (auch *has-a*-Relation) ist, wie auch die Hypernymrelation, eine Relation zwischen Synsets. Die formale Definition lautet wie folgt (Miller 1993):

*X ist ein Meronym von Y, wenn X ein Teil von Y ist.*

Die Umkehrrelation zur Meronymrelation ist die Holonymrelation:

*X ist ein Holonym von Y, wenn Y ein Teil von X ist.*

Die Meronymrelation ist hierarchisch, jedoch nicht in einer baumartigen Weise wie die Hypernymrelation, da zu einem Synset häufig mehrere Holonyme gehören. In Abbildung 2.2 ist dies ersichtlich: *table* hat die Meronyme *leg*, *tabletop* und *tableware*, *leg* ist aber auch Meronym von (unter anderem) *table*, *chair* und *tripod*, hat also mehrere Holonyme.

Die Meronymierelation ist transitiv (Miller 1993), d.h. Synsets vererben mit der Meronymrelation ihre die Eigenschaft des “Teil seins”. Wenn *X* ein Teil von *Y* ist, und *Y* ein Teil von *Z*, dann ist auch *X* ein Teil von *Z*.

## 2.3 FrameNet

FrameNet ist, wie WordNet, eine strukturiertes lexikalisch-semantisches Netz für die englische Sprache, das durch einen gerichteten Graph dargestellt wird. FrameNet wird an der University of California in Berkeley entwickelt. Die FrameNet-Datenbank kann sowohl online<sup>4</sup> abgefragt werden, als auch in Form mehrerer XML-Dateien offline genutzt werden. Lemmatisierte Wörter (*lexical units*) werden in Frames organisiert, die die Grundbausteine von FrameNet darstellen.

Ein Frame ist eine idealisierte Beschreibung einer Situation, eines Objektes oder eines Ereignisses (Ruppenhofer 2006) und besteht aus verschiedenen Elementen:

**Definition** Die Frame-Definition beschreibt die Situation, das Objekt oder das Ereignis, das der Frame darstellt und wird oft durch annotierte Beispielsätze verdeutlicht.

<sup>4</sup>Web-Oberfläche für FrameNet: <http://framenet.icsi.berkeley.edu/>



**Frame-Elemente** Frame-Elemente (kurz: FEs) stellen die Rollen in einer von einem Frame beschriebenen Situation, Ereignis oder Objekt dar. Die Frame-Elemente sind ebenfalls mit Definitionen und häufig mit annotierten Beispielsätzen versehen.

**Lexikalische Einheiten** Lexikalische Einheiten (*lexical units*) sind lemmatisierte Formen von Wörtern, die einen Frame evozieren können. Dabei wird nicht zwischen Wortarten unterschieden. In einem Frame können Wörter aller Wortarten als lexikalische Einheit auftreten, die durch ein Postfix, das durch einen Punkt an das Lemma angehängt wird, gekennzeichnet wird. Wörter, die mehrere Lesarten haben, können in verschiedenen Frames auftauchen.

**Relationen** Frames sind durch Relationen mit anderen Frames verbunden. Die Relationen zwischen Frames werden in Abschnitt 2.3.1 genauer untersucht.

Beispielsweise hat der Frame *Make\_noise* unter anderem die Frame-Elemente *Noisy\_event* (ein Ereignis, das ein Geräusch verursacht), *Sound* (das Geräusch) und *Sound\_source* (die Ursache des Geräuschs) und wird evoziert durch insgesamt 101 lexikalischen Einheiten, darunter *babble.v*, *bleat.v*, *grunt.v*, *snore.v*, *yelp.v* und *yodel.v*.

FrameNet enthält in der in dieser Arbeit verwendeten Version 1.3 rund 10000 lexikalische Einheiten, die sich ungefähr auf 4200 Nomen, 4100 Verben, 1700 Adjektive und einige wenige Wörter andere Wortarten aufteilen. Diese 10000 lexikalischen Einheiten sind auf fast 800 Frames verteilt.

### 2.3.1 Relationen in FrameNet

Relationen in FrameNet haben jeweils einen dominierenden und einen dominierten Frame, den Super-Frame und den Sub-Frame.

**Inheritance** Die Vererbungsrelation ist die stärkste Relation in FrameNet (Ruppenhofer 2006) und entspricht in etwa der *is-a*-Relation anderer Ontologien. Der Super-Frame der Vererbungsrelation ist der *Parent*-Frame, der Sub-Frame ist der *Child*-Frame. Der beerbte Frame erhält alle semantischen Eigenschaften des *Parent*-Frames, inklusive aller Frame-Elemente und Frame-Relationen, und kann diese wiederum weiter vererben (Ruppenhofer 2006). In Abbildung 2.3<sup>5</sup> ist die Vererbungsrelation mit einem roten Pfeil gekennzeichnet, d.h. der Frame *Cause\_to\_make\_noise* erbt die semantischen Eigenschaften vom Frame *Transitive\_action* und der Frame *Communication\_noise* erbt die semantischen Eigenschaften von *Communication*.

**Using** Die Using-Relation ist eine abgeschwächte Fassung der Vererbungsrelation und wird genutzt, wenn ein Teilaspekt eines *Child*-Frames von einem übergeordneten Frame erfüllt wird, das heißt ein Teil des Sub-Frames in einer Using-Relation nutzt Eigenschaften des Super-Frames.

In Abbildung 2.3 wird die Using-Relation durch gestrichelte grüne Pfeile dargestellt. Der Frame *Communication\_noise* nutzt also semantische

---

<sup>5</sup>Grafik erstellt mit dem FrameNet-II-FrameGrapher (<http://framenet.icsi.berkeley.edu/FrameGrapher/>)

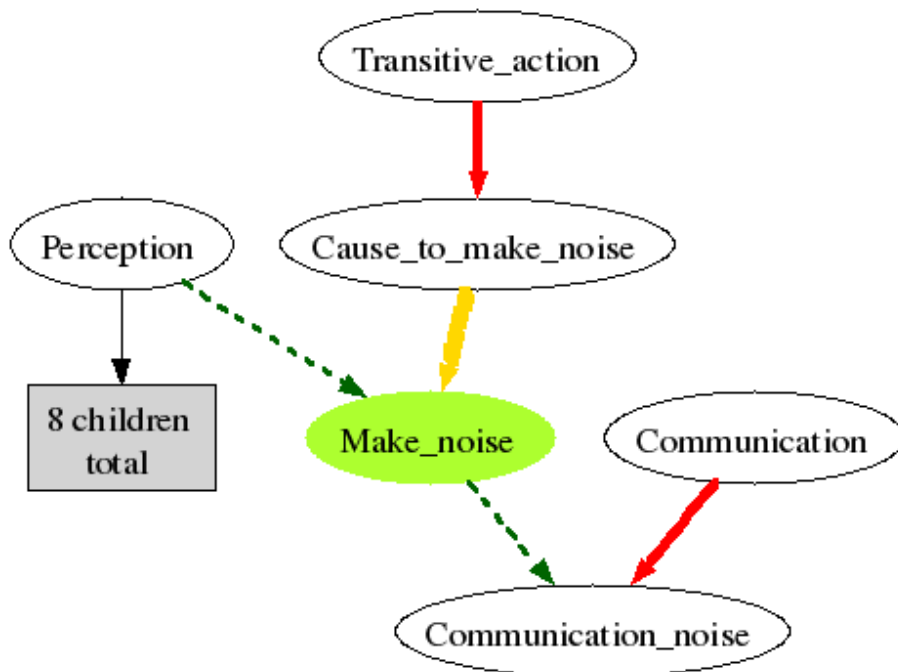


Abbildung 2.3: Frame *Make\_noise* in Relation zu anderen Frames

Eigenschaften des Frames *Make\_noise*, der wiederum semantische Eigenschaften des Frames *Perception* nutzt.

**SubFrame** Einige Frames stellen komplexe Situationen dar, so dass sie in weitere Unterframes aufgeteilt werden können. Der Super-Frame der SubFrame-Relation wird *Complex* genannt, der Sub-Frame der SubFrame-Relation ist der *Component*-Frame. Die Frame-Elemente des *Complex*-Frames können auf Frame-Elemente des *Component*-Frames abgebildet werden, es ist allerdings nicht, im Gegensatz zur Inheritance-Relation, zwingend vorgeschrieben.

In Abbildung 2.4 ist die SubFrame-Relation mit blau-gestrichelten Pfeilen dargestellt. Der Frame *Commerce\_scenario* hat die SubFrames *Having\_commercial\_agreement* und *Commercial\_transaction*, wobei letzterer wiederum ein komplexer Frame ist und die SubFrames *Commerce\_money-transfer* und *Commerce\_goods-transfer* besitzt.

**Perspective\_on** Die *Perspective\_on*-Relation wird benutzt wenn man eine Situation von mindestens zwei Standpunkten aus betrachten kann. Der Super-Frame der *Perspective\_on*-Relation ist der *Neutral*-Frame, der Sub-Frame ist der *Perspectivized*-Frame.

Der Frame *Employment\_start* beschreibt den Beginn eines Arbeitsverhältnisses. Der Beginn eines Arbeitsverhältnis kann von verschiedenen Standpunkten betrachtet werden: Vom Standpunkt desjenigen, der eingestellt wird und vom Standpunkt desjenigen, der einstellt. Diese Situationen werden

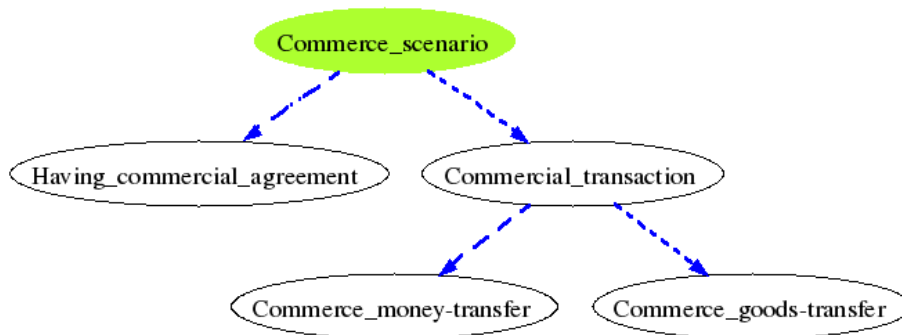


Abbildung 2.4: Frame *Commerce\_scenario* in Relation zu anderen Frames

in den Frames *Get.a.job* und *Hiring* beschrieben. Um die Frames in eine Beziehung zu setzen, verweist jeder der beiden Frames mittels einer *Perspective\_on*-Relation auf den Frame *Employment.start*.

**Precedes** Die *Precedes*-Relation wird zwischen *Component*-Frames eines komplexen Frames verwendet. Sie ordnet den Frames eine zeitliche Abfolge zu.

In Abbildung 2.5 sieht man Anhand des Frames *Crime\_scenario*, welcher zeitliche Ablauf zwischen den SubFrames herrscht: Bevor ein Prozess (Frame *Criminal.process*) stattfinden kann, muss eine Ermittlung stattgefunden haben (Frame *Criminal.investigation*), bevor eine Ermittlung stattfindet, muss ein Verbrechen begangen worden sein (Frame *Committing.crime*).

**Causative\_of/Inchoative\_of** Frames, die mit den Relationen *Causative\_of* und *Inchoative\_of* verbunden sind, treten als Frame-Tripel, bestehend aus dem *causative*-Frame, dem *inchoative*-Frame und dem *stative*-Frame, auf (Petruck et al. 2004). Der *causative*-Frame ist der Grund für ein Ereignis, das durch den *inchoative*-Frame beschrieben wird, das wiederum einen Zustand auslöst, der durch den *stative*-Frame repräsentiert wird. Anhand des Frame-Tripels *Killing*, *Death* und *Dead\_or\_alive* lassen sich die Relationen aufzeigen:

*Killing*: *John drowned Martha.*  
*Death*: *Martha died.*  
*Dead\_or\_alive*: *Martha was dead.*

Der Frame *Killing* (*causative*) ist mit einer *Causative\_of*-Relation mit dem Frame *Death* (*inchoative*) verbunden, der wiederum mit einer *Inchoative\_of*-Relation mit dem Frame *Dead\_or\_alive* (*stative*) verbunden ist.

**See.also** Die *See.also*-Relation besteht zwischen Frames, die sich ähnlich sind. So wird von einem Frame (*referring entry*) zu einem ähnlichen Frame (*main entry*) eine *See.also*-Relation eingerichtet, und im Ziel-Frame wird eine Erklärung der Unterschiede der beiden Frames gegeben.

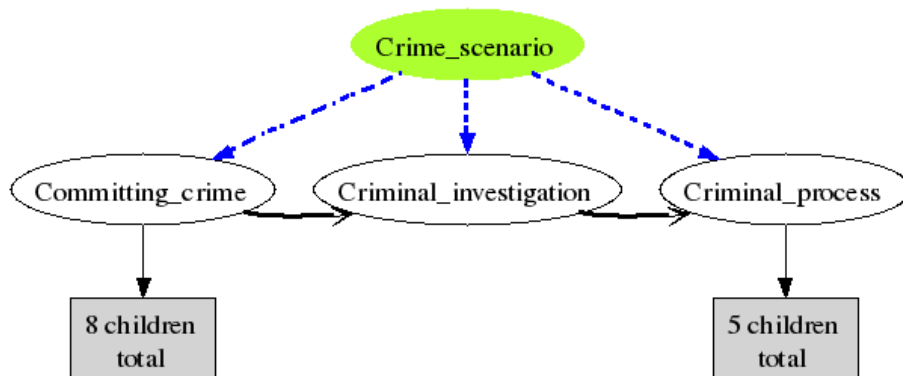


Abbildung 2.5: Frame *Crime\_scenario* mit SubFrame- und Precedes-Relationen

86 der 798 in FrameNet enthaltenen Frames stellen Wurzelknoten dar, das heißt diese Frames besitzen keine eingehenden Kanten.

## 2.4 Struktureller Vergleich

Ein großer Vorteil von WordNet ist die große lexikalische Abdeckung in Form von 206941 Lemmata, die in 117659 Synsets organisiert sind<sup>6</sup>. Demgegenüber steht das relativ kleine Lexikon von FrameNet, in dem 10195 Lemmata auf 798 Frames<sup>7</sup> verteilt sind. Dadurch ist FrameNet im Vergleich zu WordNet sehr grobmaschig: Im Schnitt sind in FrameNet 12,8 Lemmata in jedem Frame enthalten, wohingegen WordNet sehr viel feingliedriger ist und im Schnitt nur 1,76 Lemmata pro Synset enthalten sind. Diese Grobmaschigkeit von FrameNet muss natürlich zu einer Grobmaschigkeit bei den Ergebnissen von Berechnungen semantischer Ähnlichkeit führen: Semantisch sehr ähnliche Wörter sind in FrameNet mit hoher Wahrscheinlichkeit im gleichen Frame vertreten.

Die Grundbausteine von WordNet sind die Synsets: Wörter mit synonyme Bedeutung werden zu einem Synset zusammengefasst. FrameNet verfolgt einen gänzlich anderen Ansatz: Der Grundbaustein von FrameNet ist der Frame, der eine idealisierte Beschreibung einer Situation, eines Ereignisses oder einer Objekts darstellt. Einem Frame werden lexikalische Einheiten zugeordnet, die thematisch zum Frame gehören, wodurch auch synonyme Wörter im gleichen Frame zu finden sind. Bei der Zuordnung der lexikalischen Einheiten zum Frame spielt es keine Rolle, welcher Wortart die Wörter angehören – in einem Frame können alle Wortarten vertreten sein. Dies hat den Vorteil, dass semantische Ähnlichkeit mit Hilfe von FrameNet über Wortartgrenzen hinweg berechnet werden kann, wohingegen man in WordNet bei Verwendung struktureller Berechnungsmethoden auf die Berechnung semantischer Ähnlichkeit innerhalb der Wortartgrenzen beschränkt ist.

WordNet ist zudem stark auf Nomen fixiert: Rund 71% der in WordNet erfassten Lemmata sind Nomen. In FrameNet ist das Verhältnis zwischen Nomen und Verben sehr ausgeglichen: 40,8% der erfassten Wortarten sind Nomen,

<sup>6</sup>Stand: 18. Juni 2008

<sup>7</sup>FrameNet 1.3

<b>Relation</b>	<b>Anzahl Frames</b>	<b>davon mit Eingangsgrad 1</b>
Inheritance	463	384
Using	486	312
Inheritance + Using	732	475

Tabelle 2.2: Anzahl von Frames mit Eingangsgrad 1 bzgl. Inheritance und Using

<b>Relation</b>	<b>Anzahl Frames</b>
Using	486
Inheritance	463
Subframe	112
Causative_of	72
Perspective_on	63
Inchoative_of	24

Tabelle 2.3: Frame-Relationen

40,2% sind Verben.

Die für viele Methoden zur Berechnung semantischer Ähnlichkeit notwendige Hyponym-Relation ist in WordNet nur für Nomen und Verben definiert, wodurch die Anwendung dieser Methoden für Adjektive und Adverbien nicht möglich ist.

Die wichtigste lexikalische Relation in WordNet ist die Synonymie, durch die die Grundbausteine von WordNet, die Synsets, definiert sind. Die prominenteste semantische Relation zwischen Synsets ist die Hyponymierelation. WordNet wurde von Anfang an mit der Hyponymrelation als "Infrastruktur" geplant (Miller 1990), somit nimmt die Hyponymrelation eine übergeordnete Hauptrolle als semantische Relationen zwischen Synsets ein. Synsets übertragen durch die Hyponymierelation alle semantischen Eigenschaften auf ihre Kinder. Diese Hauptrelation verleiht WordNet eine Baum- bzw. Waldstruktur, da Synsets, bis auf wenige Ausnahmen, nur ein Hypernym haben. Andere Relationen können in WordNet eher als Zusatzinformationen betrachtet werden: Die eigentliche Infrastruktur von WordNet stellen die Hyponymrelationen dar, die von strukturbasierten Methoden zur Berechnung semantischer Ähnlichkeit genutzt wird.

Eine ähnlich prominente Relation wie die Hyponymrelation von WordNet existiert in FrameNet nicht. Zwar hat die Inheritance-Relation (und in geringerem Maße die Using-Relation) eine ähnliche Funktion in FrameNet, allerdings ist diese Relation nicht so stark hervorgehoben wie die Hyponymie in WordNet. Andere Relationen spielen für die FrameNet-Infrastruktur ähnlich wichtige Rollen. Auch ergibt sich keine Baumstruktur in FrameNet, wenn nur die *is-a*-ähnlichen Relationen Inheritance und Using betrachtet werden, da, trotz der Einschränkung, eine große Anzahl von Knoten einen Eingangsgrad  $> 1$  hat (siehe Tabelle 2.2).

Aus der Tabelle ist auch ersichtlich, dass nur etwas mehr als die Hälfte aller Frames von FrameNet in einer Inheritance-Relation zu einem anderen Frame stehen, wodurch die Annahme, dass die Rolle der Inheritance-Relation in FrameNet nicht so stark ist, wie die der Hyponymrelation in WordNet, bestätigt wird. In Tabelle 2.3 ist zu sehen, dass Inheritance und Using in etwa die gleiche Anzahl an Frames verbinden und keine der Relationen stark in den Vordergrund

tritt.

Man kann sehen, dass trotz gewisser struktureller Unterschiede auch Gemeinsamkeiten zwischen WordNet und FrameNet existieren, die eine Übertragung von WordNet-Methoden auf FrameNet interessant erscheinen lassen. Insbesondere bietet FrameNet gegenüber WordNet einige Vorteile für die Berechnung semantischer Ähnlichkeit, durch die sich Möglichkeiten (wie die Berechnung semantischer Ähnlichkeit über Wortartgrenzen hinaus) ergeben, die in WordNet nicht gegeben sind.

# 3 Methoden zur Berechnung semantischer Ähnlichkeit

Im Folgenden wird ein Überblick über verschiedene Methoden zur Berechnung semantischer Ähnlichkeit gegeben, die vorwiegend mit Hilfe von WordNet realisiert wurden. Grundsätzlich kann zwischen zwei Arten von Berechnungsmethoden unterschieden werden: strukturbasierte Methoden und informationsbasierte Methoden. Strukturbasierte Methoden interpretieren Informationen, die sich aus den strukturellen Eigenschaften der zugrundeliegenden Resource ableiten lassen. Informationsbasierte Methoden nutzen ebenfalls diese Eigenschaft, ziehen jedoch zusätzliche Informationsquellen, wie statistische Informationen, Definitionen und Glossen heran, um semantische Ähnlichkeit zu berechnen.

## 3.1 Strukturbasierte Methoden

Strukturbasierte Methoden zur Berechnung semantischer Ähnlichkeit nutzen die Strukturinformationen von Lexika wie WordNet zur Berechnung eines Ähnlichkeitswertes. Dabei spielt die Pfadlänge bzw. die Tiefe der Einbettung innerhalb der Struktur eine große Rolle.

Um die Berechnung der semantischen Ähnlichkeit nicht nur auf Berechnungen innerhalb der elf Nomentaxonomien zu begrenzen, sondern über die Grenzen der *unique beginners* hinaus zu ermöglichen, wurde im WordNet::Similarity-Perl-Paket (Pedersen 2004) ein artifizierlicher Wurzelknoten eingeführt, über den die elf *unique beginners* verbunden sind. Da dieses Paket zur Evaluation in dieser Arbeit benutzt wird, wird im Folgenden von der Existenz des artifizierlichen Wurzelknotens ausgegangen. Eine solche Verbindung existiert allerdings nicht zwischen den vier verschiedenen Wortarten (Nomen, Verben, Adjektive, Adverbien), die in WordNet präsent sind.

### 3.1.1 Leacock/Chodorow (1998)

Leacock und Chodorow (1998) nutzen zur Berechnung der semantischen Ähnlichkeit die normalisierte Pfadlänge zwischen zwei Konzepten (Synsets) in WordNet unter der Annahme, dass semantisch ähnliche Konzepte in der WordNet-Struktur nah beieinander liegen:

$$sim_{LC}(c1, c2) = -\log \frac{len(c1, c2)}{2 \times \max_{c \in WordNet} depth(c)}$$

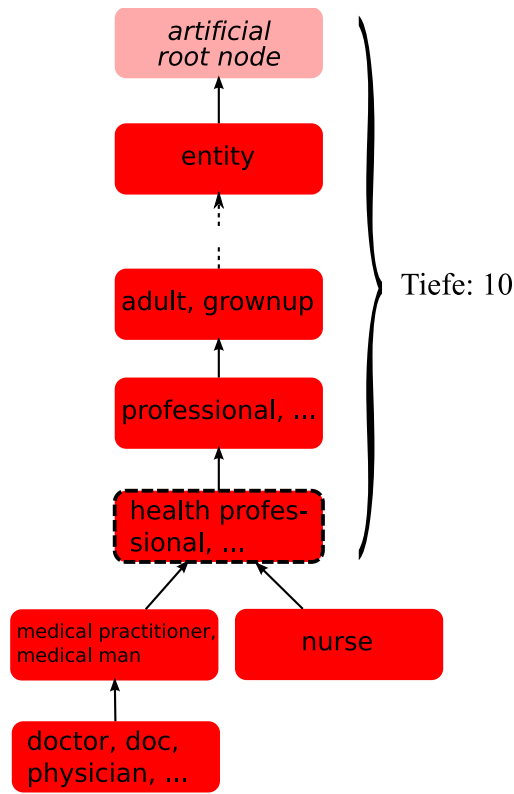


Abbildung 3.1: Teilbaum aus WordNet zur Berechnung der semantischen Ähnlichkeit zwischen *doctor* und *nurse*

$c1$  und  $c2$  stellen Konzepte in WordNet dar,  $len(c1, c2)$  die Länge des kürzesten Pfades von  $c1$  zu  $c2$ .  $c$  repräsentiert das Konzept, für das die Länge des Pfades zum künstlichen Wurzelknoten am größten ist, also  $\max_{c \in WordNet} len(c, root)$ .  $len(c, root)$  kann auch als  $depth(c)$  aufgefasst werden (Budanitsky & Hirst 2006).  $\max_{c \in WordNet} depth(c)$  repräsentiert also den Wert für die maximale Tiefe eines Knotens in WordNet. Für die Nomentaxonomie von WordNet beträgt die maximale Tiefe 16 (Budanitsky & Hirst 2006). Leacock und Chodorow messen die Länge des Pfades zwischen den Konzepten  $c1$  und  $c2$  und normalisieren den Wert durch die maximale Tiefe in der Taxonomie, in der  $c1$  und  $c2$  enthalten sind. Zu beachten ist hierbei, dass die Pfadlänge nicht anhand gemachter “Schritte” im Graphen gezählt, sondern dass die Anzahl der besuchten Knoten gezählt wird. Das heißt, wenn  $c1 = c2$  gilt, beträgt die Pfadlänge 1 und nicht 0.

### 3.1.2 Wu/Palmer (1994)

Die Methode von Wu und Palmer (1994) zur Berechnung einer “conceptual similarity” basiert auf der Baumartigkeit von WordNet, die gegeben ist, wenn nur die *is-a*-Relationen Hyponymie und Hypernymie betrachtet werden:

$$sim_{WP}(c1, c2) = \frac{2 \times depth(lso(c1, c2))}{len(c1, lso(c1, c2)) + len(c2, lso(c1, c2)) + 2 \times depth(lso(c1, c2))}$$



Zunächst wird die Tiefe des *lowest super-ordinate* (*lso*) von *c1* und *c2* berechnet. Der *lso* stellt den Knoten dar, der das Konzept repräsentiert, das die semantischen Eigenschaften der beiden Konzepte *c1* und *c2* subsumiert. Die Tiefe dieses Knotens ist die Länge des Pfades von artifiziiellen Wurzelknoten bis zu diesem Knoten ( $depth(lso(c1, c2))$ ). Skaliert wird dieser Wert durch die Summe der Pfadlängen von *c1* zum *lso* ( $len(c1, lso(c1, c2))$ ) und von *c2* zum *lso* ( $len(c2, lso(c1, c2))$ ) und der Tiefe des *lso*s. Auch bei Wu und Palmer wird die Länge der Pfade über die Anzahl der besuchten Knoten statt über die Anzahl der gemachten Schritte ermittelt (Wu & Palmer 1994).

Abbildung 3.1 zeigt anhand des Beispiels *doctor* und *nurse* die Vorgehensweise der Methode von Wu und Palmer. Wenn mittels der Hypernymrelation in der Baumhierarchie aufwärts gewandert wird, findet sich der *lowest super-ordinate* an dem Knoten, an dem sich die aufwärts gerichteten Pfade, die von den Konzepten *c1* und *c2* ausgehen, treffen. Der so gefundene Knoten vereint die semantischen Eigenschaften von *doctor* und *nurse*: das Konzept *health professional*. *health professional* hat eine Tiefe von 10 in der baumartigen WordNet-Hierarchie, es werden also von *health professional* bis zum artifiziiellen Wurzelknoten 10 Knoten besucht. Die Länge des Pfades von *doctor* zum *lowest super-ordinate* beträgt 3, die Länge des Pfades von *nurse* zum *lowest super-ordinate* beträgt 2. Für *nurse* und *doctor* errechnet sich also ein relativ hoher semantischer Ähnlichkeitswert von 0,8<sup>1</sup>.

### 3.1.3 Unterschiede zwischen Leacock/Chodorow und Wu/Palmer

Sowohl Leacock und Chodorow als auch Wu und Palmer nutzen die Pfadlänge zwischen zwei Konzepten in WordNet, um semantische Ähnlichkeit zu berechnen. Die Methode von Wu und Palmer greift jedoch stark auf die Baumhaftigkeit von WordNet zurück, die gegeben ist, wenn nur Hyper- und Hyponomie-Relationen betrachtet werden. Sind bei Leacock und Chodorow durchaus mehrere Richtungswechsel im Pfad möglich, so ist der Richtungswechsel bei Wu und Palmer nur ein einziges Mal erlaubt: an der *lowest super-ordinate*.

Die Methode von Wu/Palmer berücksichtigt im Gegensatz zur Methode von Leacock/Chodorow die Tiefe der Einbettung der Konzepte in der Hierarchie, indem sie die Tiefeninformation des *lowest super-ordinate* in die Berechnung mit einbezieht. Dadurch ist es möglich speziellen Konzepten höhere Ähnlichkeitswerte zuzuweisen als allgemeinen Konzepten, die in der Hierarchie weiter oben angesiedelt sind.

## 3.2 Informationsbasierte Methoden

Informationsbasierte Methoden zur Berechnung semantischer Ähnlichkeit bedienen sich, wie die strukturbasierten Methoden, der semantischen Informationen die aus der Struktur der Ressource gewonnen werden können, ziehen aber zusätzlich Informationen aus externen Ressourcen hinzu. Diese externen Ressourcen können beispielsweise Korpora sein, mit deren Hilfe sich statistische Informationen über Worthäufigkeiten errechnen lassen. Einige informationsbasierte Methoden verzichten auf die Strukturinformationen der Ressource und

<sup>1</sup>Der maximale Ähnlichkeitswert für die Methode von Wu/Palmer beträgt 1,0.

verwenden stattdessen Definitionen oder Glossen von Konzepten zur Berechnung semantischer Ähnlichkeit.

### 3.2.1 Resnik (1995)

Nach Resnik ist ein Kriterium, das semantische Ähnlichkeit zwischen zwei Konzepten anzeigt, die Schnittmenge der Informationen, die in den Konzepten enthalten sind. Resniks Ansatz ist ähnlich dem von Wu und Palmer: Auch Resnik sucht eine *lowest super-ordinate* in einer *is-a*-Hierarchie in WordNet, also ein Konzept, das die semantischen Eigenschaften von zwei Konzepten subsumiert. Dabei geht er davon aus, dass, je weiter man in der Struktur nach oben wandern muss, desto allgemeiner ist das Konzept, das den *lowest super-ordinate* darstellt, desto weniger semantisch ähnlich sind die zu betrachtenden Konzepte.

Resnik arbeitet hierbei mit Wahrscheinlichkeiten:  $p(c)$  sei die Wahrscheinlichkeit, in einem Text auf eine Instanz des Konzepts  $c$  zu treffen.

Nach Resnik kann die semantische Ähnlichkeit zweier Konzepte  $c_1$  und  $c_2$  durch den Informationsgehalt<sup>2</sup> ihres *lowest super-ordinate* ausgedrückt werden. Der Informationsgehalt eines Konzepts  $IC(c)$  lässt sich wie folgt berechnen:

$$IC(c) = -\log p(c)$$

Somit lässt sich die Ähnlichkeit zwischen zwei Konzepten  $c_1$  und  $c_2$  folgendermaßen berechnen:

$$sim_R(c_1, c_2) = -\log p(lso(c_1, c_2))$$

Wie kann diese Wahrscheinlichkeit berechnet werden? Resnik nutzt dazu den Brown Corpus of American English und zählt die Vorkommen der Nomen im Korpus. Je größer die Anzahl der Wortvorkommen im Korpus, desto höher ist die Wahrscheinlichkeit auf ein Konzept zu treffen, in dem das Wort enthalten ist. Dabei erhöht jedes gezählte Vorkommen einer Instanz eines Konzepts  $c_1$  auch gleichzeitig den Zähler aller übergeordneten Konzepte von  $c_1$  in der *is-a*-Hierarchie. Es gilt also  $p(c_1) \leq p(c_2)$ , wenn  $c_1$  in einer *is-a*-Hierarchie tiefer als  $c_2$  liegt.

Der Wahrscheinlichkeitswert für ein Konzept  $c$  wird also wie folgt berechnet:

$$p(c) = \frac{count(c)}{n}$$

$count(c)$  stellt dabei die gezählte Anzahl der Instanzen von  $c$  im Korpus dar, sowie die Vorkommen aller Instanzen der Konzepte, die in der *is-a*-Hierarchie unter  $c$  liegen.  $n$  ist die Gesamtanzahl aller gezählten Instanzen.

Eingesetzt in die obige Formel zur Berechnung der semantischen Ähnlichkeit, ergibt sich folgende Formel:

$$sim_R(c_1, c_2) = -\log \frac{count(lso(c_1, c_2))}{n}$$

In der Nomentaxonomie von WordNet weist der artifizielle Wurzelknoten den höchsten Zählerstand auf, da sein Zähler mit jedem im Korpus gefundenen Nomen, das in WordNet enthalten ist, erhöht wird. Der Zählerstand beträgt also  $n$ , und somit erhält der artifizielle Wurzelknoten die Wahrscheinlichkeit 1.

Für den Fall, dass zwei Konzepte  $c_1$  und  $c_2$  nur über den artifiziellen Wurzelknoten, das allgemeinst mögliche Konzept, verbunden sind, ergibt sich nach obiger Formel also eine Ähnlichkeit von  $-\log \frac{n}{n} = -\log 1 = 0$ .

<sup>2</sup><http://de.wikipedia.org/wiki/Informationsgehalt>, Stand: 29. Juni 2008

### 3.2.2 Lin (1998)

Lin entwickelte ein informationsbasiertes Ähnlichkeitsmaß, das sowohl domänen- als auch ressourcenunabhängig funktioniert. Es basiert auf drei Hypothesen, aus denen sich Regeln zur Berechnung von selbst ergeben:

1. Die Ähnlichkeit zweier Objekte  $A$  und  $B$  steht in Beziehung zu ihren Gemeinsamkeiten: Je mehr Gemeinsamkeiten sie haben, desto ähnlicher sind sie.
2. Die Ähnlichkeit zweier Objekte  $A$  und  $B$  steht in Beziehung zu ihren Unterschieden: Je mehr Unterschiede sie haben, desto weniger ähnlich sind sie.
3. Die größte Ähnlichkeit zwischen  $A$  und  $B$  ist gegeben, wenn  $A$  und  $B$  identisch sind.

Die Gemeinsamkeiten von  $A$  und  $B$  definiert Lin als den Informationsgehalt der Aussage, die die Gemeinsamkeiten von  $A$  und  $B$  beschreibt ( $-\log comm(A, B)$ ), die Unterschiede von  $A$  und  $B$  als die Differenz zwischen dem Informationsgehalt der Aussagen, die  $A$  und  $B$  beschreiben ( $-\log descr(A, B)$ ) und dem Informationsgehalt der Aussage, die die Gemeinsamkeiten von  $A$  und  $B$  beschreibt:

$$(-\log comm(A, B)) - (-\log descr(A, B))$$

Lin stellt nun folgendes Theorem auf und beweist es:

**Ähnlichkeitstheorem** Die Ähnlichkeit von  $A$  und  $B$  wird durch das Verhältnis des Informationsgehalts der Aussage, die die Gemeinsamkeiten von  $A$  und  $B$  beschreibt und dem Informationsgehalt der Aussagen, die  $A$  und  $B$  vollständig beschreiben, bemessen:

$$sim_L(A, B) = \frac{\log p(comm(A, B))}{\log p(descr(A, B))}$$

Nach Lin lässt sich daraus für Taxonomien wie WordNet folgern:

$$sim_L(A, B) = \frac{2 \times \log p(iso(c_1, c_2))}{\log p(c_1) + \log p(c_2)}$$

Die Wahrscheinlichkeiten können analog zum Vorgehen von Resnik mit Hilfe eines Korpus berechnet werden.

# 4 Anwendung der Methoden auf FrameNet

In diesem Kapitel wird die Möglichkeit der Anwendung von WordNet-Methoden zur Berechnung semantischer Ähnlichkeit auf FrameNet untersucht. FrameNet und WordNet verfolgen unterschiedliche Ansätze eines semantisch-lexikalischen Netzes, bieten jedoch genügend strukturelle Gemeinsamkeiten, die eine Umsetzung von Methoden, die in WordNet erfolgreich Anwendung finden, sinnvoll erscheinen lassen.

Für den Versuch WordNet-Methoden auf FrameNet zu übertragen, wurden die strukturbasierten Methoden von Leacock/Chodorow und Wu/Palmer gewählt. Informationsbasierte Methoden scheinen für erste Übertragungsversuche nicht geeignet, da sie zusätzlich auf externe Informationsquellen zugreifen und somit ein direkter Vergleich der gleichen Methoden in Anwendung auf WordNet bzw. FrameNet durch externe Faktoren beeinflusst würde. Durch die Wahl der strukturbasierten Methoden wird gewährleistet, dass ein Ähnlichkeitsmaß nur aufgrund der Informationen, die in der Struktur von WordNet bzw. FrameNet enkodiert sind, errechnet wird.

Leacock/Chodorow und Wu/Palmer verfolgen unterschiedliche Ansätze: Leacock und Chodorows Methode basiert auf der normalisierten Pfadlänge zwischen zwei Konzepten, wohingegen Wu und Palmer die Baumhaftigkeit von WordNet in Bezug auf *is-a*-Relationen nutzen und den in der Hierarchie am tiefsten liegenden Knoten suchen, der zwei Konzepte subsumiert: den *lowest super-ordinate*.

## 4.1 Anwendung von Leacock/Chodorow auf FrameNet

Rufen wir uns Leacock und Chodorows Methode noch einmal in Erinnerung:

$$sim_{LC}(c1, c2) = -\log \frac{len(c1, c2)}{2 \times \max_{c \in WordNet} depth(c)}$$

Die Methode berechnet die semantische Ähnlichkeit mit Hilfe einer normalisierten Pfadlänge unter der Annahme, dass semantisch ähnliche Konzepte in der WordNet-Struktur nah beieinander liegen. Um diese Methode auf FrameNet zu übertragen, müssen wir uns zunächst klar machen, welche FrameNet-Relationen für die Berechnung semantischer Ähnlichkeit von Bedeutung sein können.

In der WordNet-Variante wird mit Hilfe der Hyponomie (*is-a*-Relationen) die semantische Ähnlichkeit berechnet. Die *is-a*-Relation setzt voraus, dass alle semantischen Eigenschaften eines Konzepts auf seine Kinder vererbt werden. Die Entsprechung hierfür in FrameNet findet sich in der Inheritance-Relation, die alle semantischen Eigenschaften, sowie alle FEs und Relationen eines Frames auf seine Kinder vererbt. Die Vererbung stellt die stärkste und aussagekräftigste Relation in FrameNet dar.

Die gegebenen Voraussetzungen für die *is-a*-Relation werden zum Teil auch von der Using-Relation erfüllt, bei deren Anwendung eine Teilvererbung erfolgen kann. Sie kann als abgeschwächte Variante der Inheritance-Relation angesehen werden, das heißt die Using-Relation sollte in der FrameNet-Variante der Leacock/Chodorow-Methode Beachtung finden, jedoch sollte die Inheritance-Relation bevorzugt werden.

FrameNet bietet neben der Inheritance- und Using-Relation weitere Relationen (Subframe, Perspective\_on, Causative\_of, Inchoative\_of, See\_also), die semantische Informationen kodieren. Allerdings entsprechen alle diese Relationen in ihrer Funktion in keinsten Weise der Hyponomierelation von WordNet, so dass sie in der Berechnung der semantischen Ähnlichkeit in FrameNet mittels WordNet-Methoden keine Berücksichtigung finden sollten. Da durch die Beschränkung auf Inheritance- und Using-Relation die Bildung von mehreren FrameNet-Komponenten nicht zu verhindern ist, werden die weiteren FrameNet-Relationen dennoch als eine Art Sicherheitsnetz, ein Fall-Back, in der Berechnung zugelassen, so dass gewährleistet ist, dass immer ein Pfad zwischen zwei Frames gefunden werden kann, auch wenn diese in zwei Komponenten liegen sollten, die nicht durch eine Inheritance- oder Using-Relation verbunden sind. Sollte im Pfad eine oder mehrere der vier Fall-Back-Relationen genutzt werden, deutet dies darauf hin, dass nur geringe bis gar keine semantische Ähnlichkeit vorliegt.

Um den kürzesten Pfad zwischen zwei Frames zu finden, wurde der Dijkstra-Algorithmus zur Berechnung des kürzesten Pfades in gewichteten Graphen (Dijkstra 1959) implementiert und auf FrameNet angewandt.

Um die Relationen entsprechend ihrer Aussagekraft für die Berechnung der semantischen Ähnlichkeit zu markieren, wurden global in FrameNet Gewichte für die Relationen verteilt, d.h. Relationen, die bevorzugt zum Finden eines Pfades genutzt werden sollen (Inheritance, Using), erhielten ein niedrigeres Gewicht als die Relationen, die nur als Fall-Back dienen (Subframe, Perspective\_on, Causative\_of, Inchoative\_of, See\_also). Bei der Berechnung des kürzesten Pfades durch den Dijkstra-Algorithmus werden die Gewichte der Relationen, die im gefundenen Pfad genutzt werden, addiert. Es wird also eine gewichtete Pfadlänge zurückgegeben.

Die Verteilung der Gewichte wurde wie folgt gewählt: Inheritance erhielt das niedrigste Gewicht (0,7), da die Inheritance-Relation die aussagekräftigste und somit bevorzugte Relation ist. Using, als schwächere Variante der Inheritance-Relation erhielt ein etwas höheres Gewicht (0,8). Alle anderen Relationen erhielten ein Gewicht von 1,0. Diese Werte für die Gewichte haben sich am zweckdienlichsten gezeigt, nachdem mit verschiedenen Werten experimentiert wurde.

Um Leacock und Chodorows Methode auf FrameNet anzuwenden, war es nötig, die maximale Tiefe von FrameNet zu bestimmen. Da FrameNet nicht annähernd eine baumartige Struktur aufweist, wie es in WordNet der Fall ist, wenn nur die *is-a*-Relationen betrachtet wird, ist eine direkte Bestimmung der

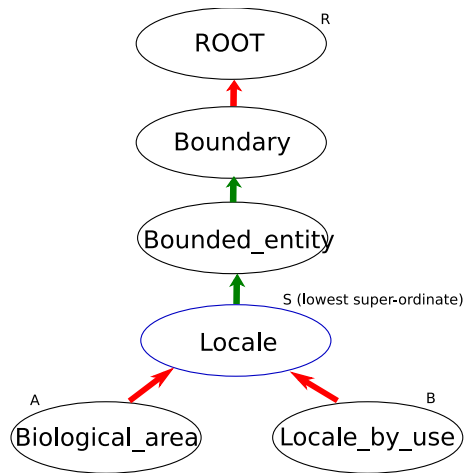


Abbildung 4.1: Teilbaum aus FrameNet zur Berechnung der semantischen Ähnlichkeit zwischen den Frames *Locale\_by\_use* und *Biological\_area*

maximalen Tiefe von FrameNet nicht möglich. Die Tiefe eines Baumes stellt jedoch nichts anderes als die Länge des Pfades von der Wurzel des Baumes (ein Knoten mit Eingangsgrad 0) zum tiefsten Blatt des Baumes (ein Knoten mit Ausgangsgrad 0) dar – es ist also der längstmögliche Pfad im Baum. Einen längstmöglichen Pfad lässt sich auch in Graphen ermitteln, mit der Einschränkung, dass der Pfad azyklisch sein muss und dass keine “Umwege” im Graph gelaufen werden dürfen: Es wird also der längstmögliche, kürzeste, azyklische Pfad im FrameNet-Graphen gesucht:

$$path_{max}(a, b) = \max_{a, b} (min(pathlen(a, b)))$$

Zu diesem Zweck wurde für alle paarweisen Kombinationen von Frames mittels des Dijkstra-Algorithmus der kürzeste Pfad zwischen den Frames bestimmt. Der längste dieser kürzesten Pfade hat eine Länge von 14 Schritten und verläuft zwischen den Frames *Connecting\_architecture* und *Behind\_the\_scenes*.

Somit waren alle Voraussetzungen erfüllt, um die Methode von Leacock und Chodorow in FrameNet anzuwenden.

## 4.2 Anwendung von Wu/Palmer auf FrameNet

Wu und Palmers Methode basiert im Gegensatz zur Methode von Leacock und Chodorow stark auf der Baumhaftigkeit, die sich in WordNet ergibt, wenn nur die Hyponymrelation in Betracht gezogen wird. Es wird versucht ein Konzept, den *lowest super-ordinate*, zu finden, das die beiden zu betrachtenden Konzepte subsumiert:

$$sim_{WP}(c1, c2) = \frac{2 \times depth(lso(c1, c2))}{len(c1, lso(c1, c2)) + len(c2, lso(c1, c2)) + 2 \times depth(lso(c1, c2))}$$

Wie in Kapitel 2.3 bereits erwähnt, sind 86 der 798 Frames von FrameNet 1.3 Wurzel-Frames, das heißt sie besitzen keinerlei eingehende Relationen.

Dies kann dazu führen, dass zwischen Frames, die unter verschiedenen Wurzeln sitzen, kein *lowest super-ordinate* gefunden werden kann. Um dieser Tatsache entgegenzuwirken, wurde ein artifizierter Wurzelknoten in FrameNet eingeführt, der mittels der Inheritance-Relation von allen 86 Wurzelknoten beerbt wird und diese somit miteinander verbindet. Dadurch wird gewährleistet, dass für jedes Frame-Paar in FrameNet ein *lowest super-ordinate* existiert. Das Konzept des artifizierten Wurzelknotens wurde in WordNet bereits genutzt, um die 11 Nomen-Taxonomien zu verbinden (siehe Kapitel 2.2).

Bezüglich der FrameNet-Relationen kann die These aus Kapitel 4.1 übernommen werden: Die *is-a*-Relationen, die ihre Entsprechung in FrameNet in der Inheritance-Relation und in abgeschwächter Form in der Using-Relation finden, sollen bevorzugt behandelt werden. Und wiederum sollen die anderen Relationen als Fall-Back beibehalten, allerdings aufgrund ihrer für die Berechnung semantischer Ähnlichkeit nicht relevanten Aussagekraft mit hohen Gewichten beschlagen werden.

Ein Problem stellt die fehlende Baumhaftigkeit von FrameNet dar, die, im Gegensatz zu WordNet, auch nicht durch alleinige Betrachtung der bevorzugten *is-a*-ähnlichen Relationen Inheritance und Using gewonnen werden kann. FrameNet wird dadurch zwar baumähnlicher, es sind aber auch bei Einschränkung auf die Inheritance- und Using-Relationen noch gut ein Drittel der Frames mit zwei oder mehr eingehenden Kanten versehen.

Das Problem lässt sich dadurch lösen, dass zunächst die Wegerichtung der Relationen eingeschränkt wird: Im FrameNet-Graph dürfen nur “nach oben”, also in Richtung des artifizierten Wurzelknotens zeigende Relationen genutzt werden. Bei der Inheritance-Relation entspricht dies der Inherits\_From-Relation – die Umkehrrelation Is\_Inherited\_By wird ausgeschlossen. Ebenso wird mit der Using-Relation verfahren: Uses ist zugelassen, Is\_Used\_By wird ausgeschlossen. Auch die Fall-Back-Relationen werden ihrer nach unten zeigenden Komponenten entledigt.

Um einen *lowest super-ordinate* zweier Frames  $A$  und  $B$  zu finden, muss ein Frame  $S$  gefunden werden, der beide Frames subsumiert und einen möglichst großen Abstand vom artifizierten Wurzelknoten  $R$  hat, also tief in der FrameNet-Hierarchie eingebettet ist. Um den *lowest super-ordinate* (Frame  $S$ ) zu finden, werden alle Pfade von Frame  $A$  und  $B$  zum artifizierten Wurzelknoten  $R$  gesucht. Dann werden alle gefundenen Pfade von Frames  $A$  und  $B$  zum Wurzelknoten betrachtet und überprüft, ob sich die Pfade teilweise überlappen. Ist dies gegeben, kann der erste Frame  $S_1$ , an dem sich die Pfade auf ihrem Weg nach oben überlappen, als ein potentieller Kandidat für den *lowest super-ordinate*-Frame  $S$  angesehen werden. So werden alle gefundenen Pfade paarweise überprüft und eine Liste von möglichen Kandidaten für den *lowest super-ordinate*  $S$  erstellt. Diese Liste wird nun abgearbeitet und die Pfadlängen der potentiellen Kandidaten zum artifizierten Wurzelknoten  $R$  gemessen. Der Frame, der am weitesten vom Wurzelknoten entfernt ist, also am tiefsten in der FrameNet-Hierarchie eingebettet ist, ist der *lowest super-ordinate*  $S$  des Frame-Paares  $A$  und  $B$  (siehe Abbildung 4.1).

Zum Auffinden der aufwärts gerichteten, kürzesten Pfade eignet sich wieder der Dijkstra-Algorithmus (Dijkstra 1959), der auf eine um die abwärts gerichteten Relationen beschnittene Version von FrameNet angewandt wird.

Anhand Abbildung 4.1 kann die Methode von Wu und Palmer nachvollzogen werden: Von `Locale.by_use` suchen wir den kürzesten, nach oben gerichteten

Pfad bis zum artifiziiellen Wurzelknoten ROOT, ebenso von Biological.area. Die Pfade sind ab dem Frame Locale identisch, somit ist Locale der gesuchte *lowest super-ordinate*, mit einer Tiefe von 3 in der Hierarchie.



# 5 Evaluation

Im Folgenden werden die auf FrameNet übertragenen Methoden auf ihre Effektivität und Funktionalität geprüft. Insbesondere wird ein Vergleich mit den WordNet-Varianten der Methoden vorgenommen.

Zur Evaluation der vorgestellten Methoden wurden zwei Ansätze gewählt: Wortpaare, die von Menschen auf semantische Ähnlichkeit hin bewertet wurden und eine in ein RTE-System eingebettete, anwendungsbezogene Evaluation.

In beiden Evaluationsverfahren zeigte sich, dass die FrameNet-Varianten der Berechnungsmethoden geringfügig bessere, in besonderen Fällen sogar erheblich bessere Ergebnisse lieferten als die WordNet-Varianten.

## 5.1 Wortpaare

Bei der Evaluation durch die Bewertung von Wortpaaren soll geprüft werden, wie nah die auf FrameNet basierenden Methoden zur Berechnung semantischer Ähnlichkeit im Vergleich zu WordNet an einen durch menschliche Bewertungen vorgegebenen Goldstandard reichen können.

Die Evaluation wurde auf einem Datenset durchgeführt, das aus der Arbeit von Rubenstein und Goodenough (1965) stammt. Rubenstein und Goodenough baten in einem Experiment 51 Kandidaten 65 Nomenpaare auf semantische Ähnlichkeit zu bewerten. Dabei sollten Punkte von 0,0 (semantisch ohne Bezug) bis 4.0 (synonym) vergeben werden.

Für die Evaluation der FrameNet-Methoden musste aufgrund der geringen lexikalischen Abdeckung von FrameNet das Datenset auf 20 der ursprünglich 65 Wortpaare eingeschränkt werden. Sollte ein Wort in mehreren Frames vertreten sein, wird, analog zu den Methoden in WordNet, zur Berechnung immer der Frame herangezogen, der den größten Ähnlichkeitswert liefert.

Für die Berechnungen der WordNet-Werte wurden WordNet 3.0 und eine Perl-Implementation verschiedener Verfahren zur Berechnung semantischer Ähnlichkeit in WordNet (Pedersen 2004) benutzt, die unter anderem auch die hier verwendeten Methoden von Leacock/Chodorow und Wu/Palmer umfasst.

Für die Berechnungen in FrameNet wurden FrameNet 1.3 und die im Rahmen dieser Arbeit entstandenen Ruby-Implementationen der Methoden von Leacock/Chodorow und Wu/Palmer benutzt.

Für den Vergleich der erhaltenen Wertereihen wurde der spearmanische Rangkorrelationskoeffizient benutzt<sup>1</sup> (Spearman 1904).

---

<sup>1</sup><http://de.wikipedia.org/wiki/Rangkorrelationskoeffizient>, Stand: 21. Juni 2008

#	Wort 1	Wort 2	Menschen	$sim_{LC}$ (WN)	$sim_{LC}$ (FN)
1	cord	smile	0,02	1,29	1,99
2	noon	string	0,04	1,21	1,80
3	automobile	wizard	0,11	1,12	1,85
4	grin	implement	0,18	1,12	1,97
5	coast	forest	0,85	1,89	2,59
6	shore	woodland	0,90	2,08	2,59
7	cemetery	woodland	1,18	1,49	2,99
8	shore	voyage	1,22	0,98	1,85
9	coast	hill	1,26	2,08	3,68
10	car	journey	1,55	0,79	2,05
11	food	fruit	2,69	1,38	5,63
12	cord	string	3,41	3,00	5,63
13	glass	tumbler	3,45	3,00	5,63
14	grin	smile	3,46	3,68	5,63
15	journey	voyage	3,58	3,00	5,63
16	coast	shore	3,60	3,00	5,63
17	forest	woodland	3,65	3,68	5,63
18	implement	tool	3,66	3,00	5,63
19	automobile	car	3,92	3,68	5,63
20	midday	noon	3,94	3,68	5,63

Tabelle 5.1: Vergleich der Leacock/Chodorow-Berechnungen für WordNet und FrameNet mit menschlichen Bewertungen

### 5.1.1 Leacock und Chodorow

In Tabelle 5.1 sind die Ergebnisse für die Berechnungen mit der Leacock/Chodorow-Methode aufgelistet, sortiert nach den menschlichen Bewertungen, beginnend mit dem Paar mit der geringsten Ähnlichkeit.

Man kann erkennen, dass die Werte für FrameNet ( $sim_{LC}$  (FN)), von wenigen Ausnahmen abgesehen, steigen oder gleichbleibend sind, wohingegen die Werte für die WordNet-Methode ( $sim_{LC}$  (WN)) häufig Ausreißer nach unten aufzeigen.

In der FrameNet-Reihe ergibt sich ab dem elften Wortpaar nur noch der Maximalwert für die Leacock/Chodorow-Berechnung in FrameNet von 5,68. Dies ist auf die Grobmaschigkeit von FrameNet zurückzuführen: Die Wörter liegen beide im gleichen Frame.

Die Berechnung der spearmanischen Rangkorrelationskoeffizienten, der Ergebniswerte zwischen -1 und 1 liefert, ergibt für FrameNet einen Wert von 0,88 und für WordNet einen Wert von 0,81. Beide sind somit recht nahe an dem Optimum von 1,0, wobei die FrameNet-Methoden sogar geringfügig besser abschneiden. Wenn die letzten zehn Paare, die durch FrameNets Grobmaschigkeit nur Maximalwerte liefern außer Acht gelassen werden, ergibt sich ein deutlich besserer Wert für FrameNet gegenüber WordNet: Die WordNet-Methode kommt auf einen Wert -0,05 und die FrameNet-Methode kommt auf 0,53 im spearmanischen Rangkorrelationskoeffizient.

#	Wort 1	Wort 2	Menschen	$sim_{WP}$ (WN)	$sim_{WP}$ (FN)
1	cord	smile	0,02	0,38	0,00
2	noon	string	0,04	0,35	0,00
3	automobile	wizard	0,11	0,46	0,00
4	grin	implement	0,18	0,25	0,00
5	coast	forest	0,85	0,62	0,67
6	shore	woodland	0,90	0,67	0,67
7	cemetery	woodland	1,18	0,50	0,75
8	shore	voyage	1,22	0,22	0,00
9	coast	hill	1,26	0,71	0,89
10	car	journey	1,55	0,19	0,00
11	food	fruit	2,69	0,47	1,00
12	cord	string	3,41	0,94	1,00
13	glass	tumbler	3,45	0,95	1,00
14	grin	smile	3,46	1,00	1,00
15	journey	voyage	3,58	0,96	1,00
16	coast	shore	3,60	0,92	1,00
17	forest	woodland	3,65	1,00	1,00
18	implement	tool	3,66	0,94	1,00
19	automobile	car	3,92	1,00	1,00
20	midday	noon	3,94	1,00	1,00

Tabelle 5.2: Vergleich der Wu-Palmer-Berechnungen für WordNet und FrameNet mit menschlichen Bewertungen

### 5.1.2 Wu und Palmer

In Tabelle 5.2 sind die Ergebnisse für die Berechnungen mit der Wu/Palmer-Methode aufgelistet.

Es ist wiederum zu erkennen, dass die Werte für WordNet oft Ausreißer nach unten haben. In der FrameNet-Wertereihe ( $sim_{WP}$  (FN)) hingegen ist, mit Ausreißern nach unten an Stelle 8 und 10, eine Steigerung oder ein Gleichbleiben der Werte innerhalb der ersten zehn Wortpaare zu beobachten. Ab dem elften Wortpaar wird nur noch der maximale Ähnlichkeitswert von 1,0 berechnet, da beide Wörter des Wortpaares im gleichen Frame liegen.

Beim spearmanischen Rangkorrelationskoeffizienten ergibt sich ebenfalls die gleiche Tendenz wie bei der Methode von Leacock und Chodorow: Für die FrameNet-Wertereihe errechnet sich ein Wert von 0,88, für die WordNet-Wertereihe ein Wert von 0,80. Somit ist auch bei der Wu/Palmer-Methode ein leichter Vorteil bei FrameNet zu verzeichnen. Wenn wiederum nur die ersten zehn Wortpaare betrachtet werden, hebt sich FrameNet nochmals deutlicher von der WordNet-Methode ab: Für FrameNet errechnet sich ein Wert von 0,53 und für WordNet ein Wert von -0,01 im spearmanischen Rangkorrelationskoeffizienten.

### 5.1.3 Diskussion

Es zeigt sich, dass FrameNet sowohl für die Methode von Leacock/Chodorow als auch für die Methode von Wu/Palmer Ergebnisse liefern kann, die näher am durch die menschlichen Bewertungen gegebenen Goldstandard liegen als die entsprechenden FrameNet-Varianten. Ein Nachteil ist die Grobmaschigkeit von Fra-

meNet im Vergleich zu WordNet: Wenn die zu bewertenden Wörter sich semantisch stark ähneln, kann FrameNet nur noch den maximalen Ähnlichkeitswert liefern. Dies liegt darin begründet, dass semantisch sehr ähnliche Wörter in FrameNet im gleichen Frame zu finden sind.

## 5.2 RTE

Da die Evaluation mittels Wortpaaren nur auf einem kleinen Datenset durchgeführt werden konnte, sollen die Methoden nun auf einem großen Datensatz durch Einbettung in ein bestehendes System geprüft werden. Dazu wurden die Korpora der RTE3-Challenge<sup>2</sup> (*recognizing textual entailment*) gewählt. In der RTE-Challenge sollen Paare von Text und Hypothese daraufhin bewertet werden, ob eine Hypothese aus einem Text gefolgert werden kann. Text und Hypothese bestehen gewöhnlich aus einem, selten aus mehreren Sätzen. Zunächst wird ein hinsichtlich positiver oder negativer Folgerung balanciertes, annotiertes Trainingsset von 800 Text-Hypothese-Paaren ausgegeben. Die auf dem Trainingsset entwickelten Methoden werden in der eigentlichen Challenge auf einem balancierten, annotierten Testset, das ebenfalls aus 800 Text-Hypothese-Paaren besteht, geprüft. In Trainings- und Testkorpus sind jeweils vier verschiedene Klassen von Text-Hypothese-Paaren vertreten: IR (Information Retrieval), IE (Information Extraction), QA (Question Answering) und SUM (Summarization).

### 5.2.1 Anwendung der Methoden auf RTE

Für die Evaluation im RTE-System wird angenommen, dass sich eine positive Folgerung der Hypothese aus dem Text dann ergibt, wenn die Wörter in Text und Hypothese semantisch ähnlich sind. Dies ist selbstverständlich eine vereinfachende Annahme, da hierbei etwa Negationen keinerlei Berücksichtigung finden.

Um dies umzusetzen, müssen Wortähnlichkeiten auf Satzebene übertragen werden. Li et al. (2007) hat in der RTE3-Challenge die semantische Ähnlichkeit zwischen Text und Hypothese auf Basis der semantischen Ähnlichkeit der Wörter von Text und Hypothese folgendermaßen berechnet:

Wenn  $H = \{HW_1, HW_2, \dots, HW_m\}$  die Menge von Wörtern der Hypothese und  $T = \{TW_1, TW_2, \dots, TW_n\}$  die Menge von Wörtern des Textes sind, dann lässt sich die lexikalisch-semantische Ähnlichkeit für ein Text-Hypothese-Paar  $LSS(H, T)$  wie folgt berechnen:

$$LSS(H, T) = \frac{\sum_i (\text{MAX}_j (\frac{SSim(HW_i, TW_j)}{SSim(HW_i, HW_i)}) \times IDF(HW_i))}{\sum_i IDF(HW_i)}$$

$SSim$  stellt in dieser Gleichung eine Funktion dar, die einen Wert für die semantische Ähnlichkeit von Wörtern liefert,  $IDF(w)$  ist eine Funktion, die die inverse Dokumentenfrequenz<sup>3</sup> für ein Wort  $w$  liefert. Die inverse Dokumentenfrequenz findet in dieser Evaluation keine Berücksichtigung, um die Berechnungen

<sup>2</sup>Recognizing Textual Entailment Challenge; <http://www.pascal-network.org/Challenges/RTE/>

<sup>3</sup><http://de.wikipedia.org/wiki/TF-IDF>, Stand: 22. Juni 2008

	Lexikalische Überlappung	Leacock/Chodorow	Wu/Palmer
WordNet	63,13%	63,5% (+ 0,37%)	61,88% (- 1,25%)
FrameNet	63,13%	63,75% (+ 0,62%)	64,00% (+ 0,83%)

Tabelle 5.3: Ergebnisse der Evaluation im RTE-System

auf die strukturelle Berechnung der semantische Ähnlichkeit zu fokussieren und die Ergebnisse nicht durch weitere Faktoren zu beeinflussen. Somit ergibt sich für die Berechnungen in dieser Arbeit folgende Gleichung:

$$LSS(H, T) = \frac{\sum_i (MAX_j (\frac{SSim(HW_i, TW_j)}{SSim(HW_i, HW_i)}))}{i}$$

Für die Evaluation wurde der RTE-Korpus mit Hilfe des TreeTaggers<sup>4</sup> (Schmid 1994) mit Wortarten und Lemmata versehen und anschließend eine Inhaltswortfilterung vorgenommen, so dass nur Nomen, Verben, Adjektive und Adverbien in der Berechnung Berücksichtigung finden. Für die Berechnungen wurden die Lemmata der Wortformen herangezogen.

Um einen Vergleich mit einem Base-Line-System ziehen zu können, wurde ein System ausgewählt, das auf lexikalische Überlappung prüft. Dazu wurde eine Abwandlung des Jaccard-Algorithmus<sup>5</sup> (Jaccard 1901) zur Berechnung des lexikalischen Überlappungsgrades benutzt, der auch in einem Beitrag zur RTE3-Challenge zum Einsatz kam (Burchardt et al. 2007). Es wird das Verhältnis der Anzahl der Wörter, die sowohl in der Hypothese als auch im Text vorkommen, zur Anzahl der Wörter in der Hypothese berechnet:

$$OL(H, T) = \frac{|H \cap T|}{|H|}$$

wobei  $H$  und  $T$  die Mengen der Wörter in Hypothese und Text darstellen. Die Ergebnisse für das Base-Line-System finden sich in den Tabellen 5.3 und 5.4 jeweils in der ersten Spalte.

Für die Berechnung mit Hilfe der Methoden von Leacock/Chodorow und Wu/Palmer wurde zunächst auf Wortgleichheit zwischen Hypothese und Text geprüft. Dies ist sinnvoll, da im RTE-Korpus zahlreiche Eigennamen vorkommen, zu denen selbst WordNet keine semantischen Informationen liefern kann. War eine Wortgleichheit gegeben, wurde der Wert für die maximale semantische Ähnlichkeit der jeweils angewandten Methode vergeben. War keine Gleichheit gegeben, wurde die semantische Ähnlichkeit mittels der oben genannten Gleichung berechnet. Für den Fall, dass eines der gegebenen Wörter nicht im Lexikon von WordNet oder FrameNet vorhanden war, wurde der minimale Ähnlichkeitswert vergeben.

Die Auswertung der Ergebnisse erfolgte mittels des Maschinlernsystems Weka<sup>6</sup>, unter Benutzung des Entscheidungsbaumklassifikators J48.

Bei dieser Evaluation ist zu beachten, dass die RTE-Datensets lexikalisch nur schlecht von FrameNet abgedeckt werden (im Durchschnitt 45% für Trainings- und Testset). Aus diesem Grund wurde die Evaluation nochmals mit einem balancierten Teilkorpus des RTE-Korpus wiederholt, der eine mindestens 60%ige lexikalische Abdeckung in FrameNet aufweist. Dies entspricht in etwa 50% des RTE-Korpus.

<sup>4</sup><http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>

<sup>5</sup>[http://en.wikipedia.org/wiki/Jaccard\\_index](http://en.wikipedia.org/wiki/Jaccard_index), Stand: 25. Juni 2008)

<sup>6</sup><http://www.cs.waikato.ac.nz/ml/weka/>

	Lexikalische Überlappung	Leacock/Chodorow	Wu/Palmer
WordNet	59,85%	62,09%	58,12%
FrameNet	59,85%	61,08%	63,79%

Tabelle 5.4: Ergebnisse der Evaluation im RTE-System für die Hälfte des RTE-Korpus, der eine mindestens 60%ige lexikalische Abdeckung in FrameNet aufweist

## 5.2.2 Diskussion

In Tabelle 5.3 sind die Ergebnisse der Evaluation im RTE-System zu sehen. Die Prozentzahlen stellen die hinsichtlich der Folgerung korrekt klassifizierten Text-Hypothese-Paare dar.

Man sieht, dass die Leacock/Chodorow-Methode sowohl in der WordNet- als auch in der FrameNet-Variante eine Verbesserung gegenüber dem Base-Line-System erreichen kann, wobei die FrameNet-Variante leicht besser abschneidet und 0,25% mehr der Text-Hypothesen-Paare korrekt klassifiziert.

Bei der Wu/Palmer-Methode verliert die WordNet-Variante gegenüber dem Base-Line-System 1,25% – es werden mehr Text-Hypothese-Paare fälschlicherweise mit *YES* klassifiziert, als es beim Base-Line-System der Fall ist. Die FrameNet-Methode hingegen gewinnt einen Prozentpunkt gegenüber dem Base-Line-System hinzu, klassifiziert somit über drei Prozent mehr der Text-Hypothese-Paare korrekt als die WordNet-Variante.

Tabelle 5.4 zeigt die Ergebnisse der Evaluation für den Teil des RTE-Korpus, für den eine mindestens 60%ige lexikalische Abdeckung in WordNet gegeben ist. Der Wert für das Base-Line-System beträgt hier 59,85%. Alle Methoden verlieren in dieser Evaluation gegenüber der Evaluation im gesamten RTE-Korpus. Dies ist darauf zurückzuführen, dass die anfänglich gemachte Annahme, eine positive Folgerung könne durch semantisch ähnliche Wörter in Text und Hypothese gefolgert werden, zu einem gewissen Maße korrekt ist, hier aber an ihre Grenzen stößt und mehr Text-Hypothese-Paare durch hohe semantische Ähnlichkeit fälschlicherweise als positiv (mit *YES*) klassifiziert werden.

Ein Beispiel, an dem sich der Sachverhalt der falschen Klassifizierung gut nachvollziehen lässt, ist Text-Hypothese-Paar 776 aus dem RTE3-Test-Korpus:

**Text:** *Last month, South Dakota moved to ban all abortion in the state, even in cases of rape and incest. Mississippi is considering similar legislation.*

**Hypothese:** *In South Dakota abortion is allowed in cases of rape and incest.*

Bis auf *allow* sind alle Inhaltswörter der Hypothese im Text wiederzufinden. Zusätzlich kann zwischen den Verben *ban* (Text) und *allow* (Hypothese) eine starke semantische Ähnlichkeit festgestellt werden (Frame Prohibitting (*ban.v*) beerbt Frame Permitting (*allow.v*)). Das Paar wird vom Maschinensystem folgerichtig mit *YES* klassifiziert, auch wenn die korrekte Klassifizierung natürlich *NO* wäre.

Ziel der Arbeit war es allerdings *nicht* ein gutes RTE-System zu entwickeln, sondern semantische Ähnlichkeit zuverlässig zu berechnen. Dies ist auch in dieser Evaluation gegeben und wird auch gerade dadurch gezeigt, dass die Werte der RTE-Berechnungen bei höherer lexikalischer Abdeckung in FrameNet durch fälschlicherweise positiv klassifizierte Text-Hypothese-Paare abnehmen.

# 6 Fazit

## 6.1 Zusammenfassung

In dieser Arbeit wurden die Fragen untersucht, ob Methoden zur Berechnung semantischer Ähnlichkeit, die in WordNet Anwendung finden, erfolgreich auf FrameNet übertragen werden können und ob die Methoden in Anwendung auf FrameNet eine ähnlich gute oder sogar bessere Performance bieten können.

Die Übertragung der gewählten strukturbasierten Methoden von Leacock/Chodorow und Wu/Palmer auf FrameNet ist mit Hilfe von Approximationen, die durch die strukturellen Unterschieden zwischen WordNet und FrameNet nötig wurden, gelungen. Für die in WordNet zur Berechnung semantischer Ähnlichkeit genutzte Hyponymrelation wurde in der Inheritance- und Using-Relation eine Entsprechung in FrameNet gefunden.

Da die Methode von Wu/Palmer stark auf einer baumartigen Struktur aufbaut, konnte in FrameNet, das die benötigte Baumstruktur nicht bietet, ein Weg gefunden werden, die Struktur von FrameNet dahingehend zu interpretieren, dass die nötigen Berechnungen dennoch im Sinne der Methode von Wu/Palmer durchgeführt werden konnten.

Die Methode von Leacock/Chodorow konnte ebenfalls in ihrem ursprünglichen Sinne umgesetzt werden, in dem zur Normalisierung der Pfadlänge, die in der WordNet-Variante durch die maximale Tiefe der WordNet-Taxonomie erfolgt, die Länge des kürzesten Pfades, der zwischen den beiden am weitesten voneinander entferntesten Frames in FrameNet gefunden werden konnte, genutzt wurde.

Durch die Evaluation mittels Wortpaaren und durch Einbettung in ein RTE-System wurde gezeigt, dass die FrameNet-Varianten eine bessere Performance als die WordNet-Varianten erreichen können, die Möglichkeiten jedoch durch das kleine Lexikon und die Grobmaschigkeit von FrameNet im Vergleich zu WordNet stark eingeschränkt sind.

## 6.2 Ausblick

Eine Möglichkeit die Ergebnisse der Berechnungen zu verfeinern, wäre, zusätzlich zur Verteilung der globalen Gewichte auf verschiedene Relationstypen, die Verteilung von lokalen Gewichten. Geschwister-Frames, die tief in der Hierarchie angesiedelt sind, sind sich ähnlicher als Geschwister-Frames, die nur sehr allgemeine Konzepte darstellen. Wenn Frames, die tief in der FrameNet-Hierarchie sitzen und somit sehr speziellere Konzepte darstellen, mit einem geringeren Ge-

wicht mit ihren Parent-Frames verbunden sind, kann dadurch eine detailliertere Berechnung der semantischen Ähnlichkeit erreicht werden.

Die Tiefeninformation wird zwar von der Wu/Palmer-Methode durch die Messung der Tiefe des *lowest super-ordinate* in der Hierarchie schon berücksichtigt, jedoch könnten auch bei dieser Methode lokale Gewichte eine Verfeinerung der Ergebnisse bewirken.

Eine Auswertung des Abschneidens der FrameNet-Methoden über die verschiedenen RTE-Klassen (Information Retrieval, Information Extraction, Summarization und Question Answering) könnte sicher interessante Denkanstöße für mögliche weitere Anwendungsszenarien der FrameNet-Methoden liefern.

Die in dieser Arbeit vorgestellten informationsbasierten Methoden zur Berechnung semantischer Ähnlichkeit bieten ebenfalls interessante Möglichkeiten einer Übertragung auf FrameNet. Insbesondere könnten dabei die Definitionen von Frames, so wie die Frame-Elemente und deren Definitionen, in die Berechnung mit einbezogen werden.



## 7 Bibliographie

Budanitsky, Alexander, Graeme Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. In: *Computational Linguistics*, 32(1), S. 13-47.

Burchardt, Aljoscha, Nils Reiter, Stefan Thater, Anette Frank. 2007. A Semantic Approach To Textual Entailment: System Evaluation and Task Analysis. In: *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prag, Tschechei, S. 10-15.

Dijkstra, E. W.. 1959. A note on two problems in connexion with graphs. In: *Numerische Mathematik*. 1, S. 269–271.

Fellbaum, Christiane, editor. 1998. WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, MA.

Hayes, Brian. 1999. The Web of Words. In: *American Scientist*, Volume 87, Number 2, S. 108.

Jaccard, Paul. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. In: *Bulletin del la Société Vaudoise des Sciences Naturelles* 37, S. 547-579.

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

Li, Baoli, Joseph Irwin, Ernest V. Garcia, Ashwin Ram. 2007. Machine Learning Based Semantic Inference: Experiments and Observations at RTE-3. In: *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prag, Tschechei, S. 159-164.

Lin, Dekang. 1998. An Information-Theoretic Definition of Similarity.

Miller, George A. 1990. Nouns in WordNet: a lexical inheritance system. In: *International Journal of Lexicography* 3 (4), S. 245-264.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and

- Katherine Miller. 1993. Introduction to WordNet: An On-line Lexical Database.
- Pedersen, Ted, Siddharth Patwardhan, Jason Michelizzi. 2004. WordNet::Similarity - Measuring the Relatedness of Concepts. In: *Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-04)*, S. 38-41, Boston, MA.
- Petruck, Miriam R. L., Charles J. Fillmore, Collin F. Baker, Michael Ellsworth, Josef Ruppenhofer. 2004. Reframing FrameNet Data. In: *Proceedings of The 11th EURALEX International Congress*, Lorient, Frankreich, S. 405-416.
- Resnik, Philip. 1995. Using information content to evaluate semantic similarity. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, S. 448-453, Montreal, Kanada.
- Rubenstein, Herbert and John B. Goodenough. 1965. Contextual correlates of synonymy. In: *Communications of the ACM*, 8(10), S. 627-633.
- Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, and Christopher R. Johnson. 2006. FrameNet II: Extended Theory and Practice. ICSI Technical Report.
- Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of International Conference on New Methods in Language Processing*.
- Spearman, Charles. 1904. The proof and measurement of association between two things. In: *Amer. J. Psychol.*, 15, S. 72-101.
- Wu, Zhibiao and Martha Palmer. 1994. Verb semantics and lexical selection. In: *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM, S. 133-138.

# Versicherung

Hiermit versichere ich, dass ich die Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

(Sven Arweiler)