Semantic Relations and User Interests

WordNet Lexical Database und mögliche Anwendungen bei der Benutzerinteraktion

Sebastian Blohm

Sorry...

Wie kann eine Maschine beim Suchen helfen?

Was Maschinen schon können:

> Schnell viel Material sichten.

> Zuverlässig Ergebnisse liefern.

> Arbeiten, ohne zu ermüden.

Der Traum: Die Maschine „versteht“, was ich brauche.

Probleme

> Unterspezifiziertheit („Mir ist kalt.“)

> Mehrdeutigkeit (Thema „Bank“)

Aspekte von Pragmatik

> User
> Umgebung
> Zweck

Identifikation und Beschreibung des **Users** ist Kernidee von Meinetz.

Die **Umgebung** ist hier ohnehin eine Suchumgebung.

Der **Zweck** der Suche kann in Form von Sprache zum Ausdruck kommen.

Semantische Analyse kann helfen, dem Zweck einer Anfrage gerecht zu werden.

Beispiele für Anwendung von Semantischem Wissen

- Interpretieren der Suchanfrage
  - Mehrdeutigkeiten ausschließen
  - Fokus eingrenzen / erweitern
- Dokumente ordnen
  - Zuordnung zu Themenbereichen klären
  - Beziehungen unter Dokumenten herstellen

Achtung: Dies sind alles nur heuristische Verfahren, keine „saubere“ Semantik.

Typen semantischer Beziehungen

> zwischen Wortformen (lexical)

Synonym: zwei Worte haben die gleiche Bedeutung.

> zwischen Wortbedeutungen (semantic)

Hyperonym, Hyponym: Oberbegriff/Spezialisierung

Antonym: Gegenteil

Meronym/Holonym: Bestandteil/ entsteht aus

Mögliche Eigenschaften dieser Beziehungen:

» transitiv

» symmetrisch oder reziprok

Die Beziehungen eigenen sich zum Aufbau einer **Taxonomie** in Graphnotation.

Beispielnyme

Xnyme des Wortes **Fußballspieler**:

Hyperonym - Sportler
Hyponym - Torwart
Meronym – Trikot
Holonym - Mannschaft
Antonym - *Angler
Synonym - ...

Symmetrie: antonym(heiß,kalt)<->antonym(kalt,heiß)
Reziprozität: hyper(Sportler,Fussballspieler)<-> hypo(Fussballspieler,Sportler)
Transitivität: hyper(Fussballspieler,Torwart) && hyper(Sportler,Fussballspieler) -> hyper(Sportler,Torwart)

Die WordNet Datenbank

WordNet baut mit diesen Beziehungen eine Lexikon-Datenbank auf.

Die Beziehungen entstehen zwischen **Synsets**: Klassen synonymer Wörter.

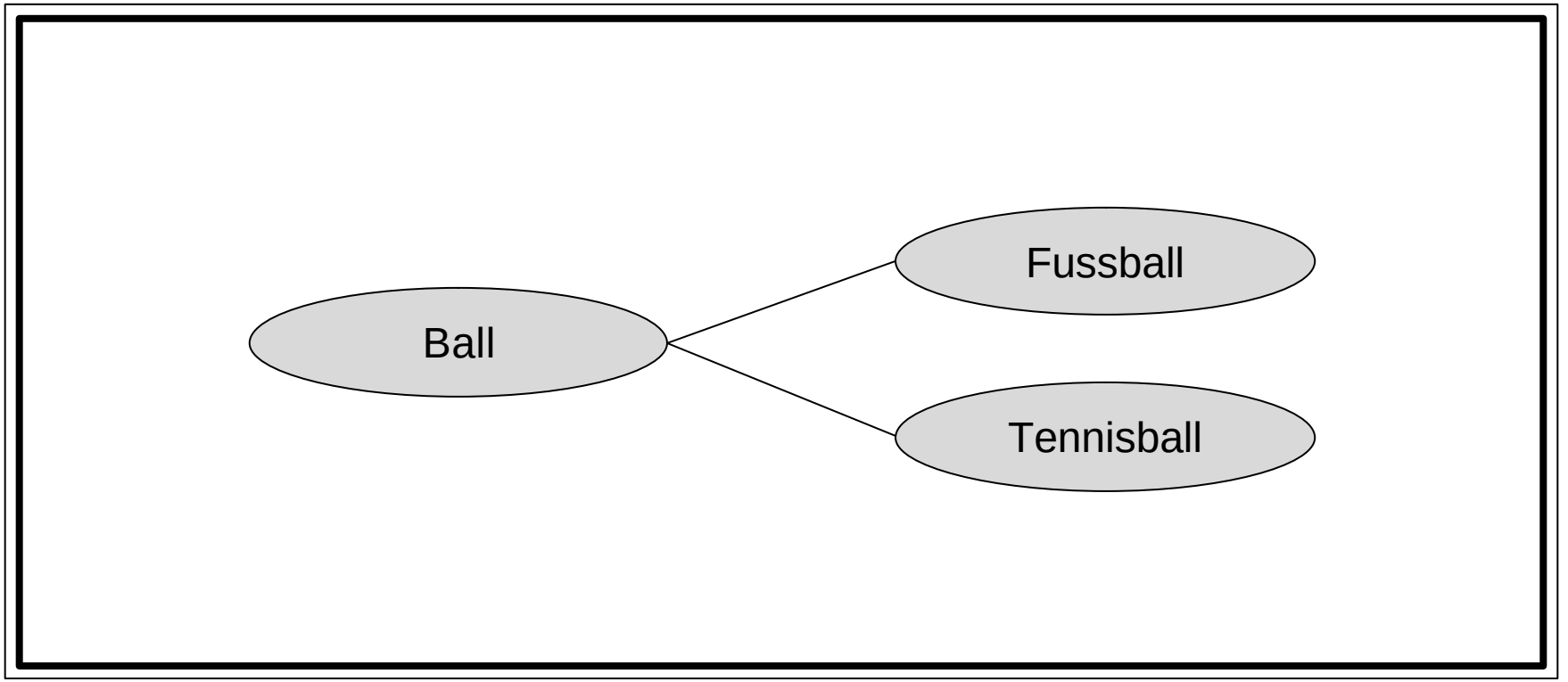
Definition der Synonymie in WordNet:

> „Two expressions are synonymuous in a linguistic context C if the substitution of one for the other in C does not alter the truth value.“

Diese Definition genügt, weil ein Wort in mehreren Synsets auftreten kann (Polysemie).

Die WordNet Datenbank - Graphen

Beispiel für graphische Repräsentation der Hyponomie-Beziehungen:



Die Wordnet – Datenbank: Browser

Zu Nomen kann der Browser Anzeigen:

> Synonyme nach Häufigkeit

> Synonyme nach Ähnlichkeit

> „Coordinate Terms“ – Alternative Begriffe des gleichen Hyperonyms

> Hyperonyme und Hyponyme

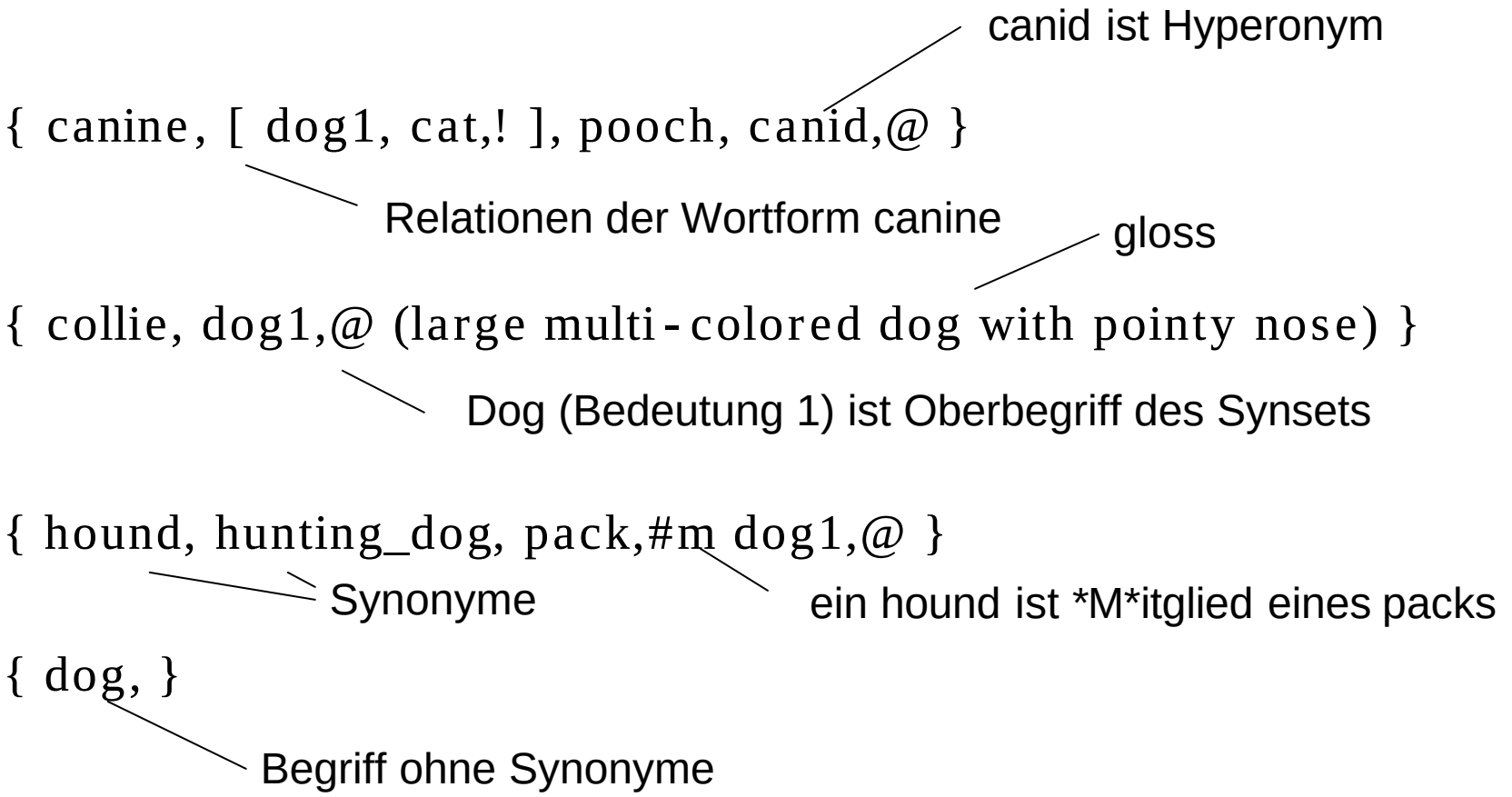
> Bekanntheit (abgeschätzt durch Polysemie)

Zu den Relationen ist auch eine für Menschen lesbare Definition (**gloss**) gespeichert.

WordNet Browser.lnk

Die Wordnet – Datenbank: Datenstruktur

Dies ist die Datenstruktur von Lexicographer-Files, die manuell erstellt werden.



WordNet und Psycholinguistik

WordNet wurde durch psycholinguistische Entdeckungen:

> Konzeptionelle Verbindungen zwischen Worten sind durch Assoziationsforschung belegt.

> Die Einteilung der Wörter in Kategorien (Nomen, Verb, usw.) entspricht u. a. Studien an Aphasiekranken.

ganz oben in WordNet

Es gibt kein leeres Synset an der Spitze von WordNet. Vielmehr existieren 25 „**unique beginners**“:

{act, activity}	**{food}**	{possession}
{animal, fauna}	{group, grouping}	{process}
{artifact}	{location}	**{quantity, amount}**
{attribute}	**{motivation, motive}**	**{relation}**
{body}	**{natural object}**	{shape}
{cognition, knowledge}	{natural phenomenon}	{state}
{communication}	**{person, human being}**	**{substance}**
{event, happening}	**{plant, flora}**	{time}
{feeling, emotion}		

Nachträglich zusammengefasst: **organism**, **object**, **abstraction**, psychological feature

Tiefe von WordNet

Die Tiefe des Hyponomie-Graphs ist „selten größer als 10 oder 12 Ebenen.“

Das ist also ein Tor:

```
goal
    => score
        => success
            => attainment
                => accomplishment, achievement
                    => action
                        => act, human action,
```

Umfang von WordNet

Natürlich ist WordNet eine Garagen-Erfolgsstory.

> Das Projekt wurde 1985 mit drei Personen gestartet.

> Es hat insgesamt ca. 3 Millionen $ US erhalten.

> Es gibt industrielle Anwendungen und internationale Folgeprojekte

> Enthält ca. 140.000 Wörter. Genauer:

Kategorie	Formen	Bedeutungen
Nomen	94474	116317
Verb	10319	22066
Adjektiv	20170	29881
Adverb	4546	5677

Grenzen von WordNet

WordNet ist hauptsächlich als (Meta-) Lexikon gedacht. Möchte man es für zum NLP nutzen ergeben sich Schwierigkeiten. Denn:

> Das zentrale Konzept der Synonymität ist recht starr definiert. Gewisse Abgrenzungen sind dadurch einfach nicht möglich.

> Es wird davon ausgegangen, dass zu jedem Konzept auch ein Wort gehört. (z.B. „Geschwister“ im Französischen)

> Kein gutes Verhältnis zur „non-monotonicity“ (Pinguin als Vogel)

> der Graph ist teilweise willkürlich gesetzt.

Grenzen von WordNet II

WordNet ist Handarbeit.

> kaum Eigennamen

Vorteile von WordNet

> effizientes Auffinden von verwandten Wörtern.
> In einer Erweiterung liegen auch statistische Daten vor.
> Glosses erlauben leichte Erfassbarkeit für User.
> Vorhandensein in Sysnsets ist ein einfaches Vergleichskriterium
> Viele Daten

Anwendungen von Wordnet (linguistische)

- **Semantic Concordances:** Begriffe in Texten werden (von Hand) semantisch eindeutig gemacht.
- **Word-Sense Disambiguation**: Distanz zwischen Wörtern= Anzahl der Nodes, die im Hyponymie-Graph dazwischen liegen. So können Umgebungswörter eingeordnet werden. (play Football ~= play Soccer)
- diverse **online Lexika**.

Semantic Concordances

- „Semantic Tagging“: Korpus Text, in dem Wörtern ihr Synset zugeordnet ist → Polysemie disambiguiert.
- Korpora: 50% der Wörter open class word.
- Polysemie in 18% der Word-Types aber 83% der Word-Token.
- Vorgang eignet sich zum Prüfen der Vollständigkeit eines Lexikons.
- Automatische Herstellung noch nicht möglich.

Welche semantischen Beziehungen sind für Suche interessant?

- Synonymie: Vor allem durch die damit implizit kodierte Polysemie (mehrere Bedeutungen eines Wortes)
  - Vermeiden mehrdeutiger Suchanfragen
  - Klassifikation von Dokumenten
- Hyponymie / Meronymie: Inferenzen auf Text.
  - ermöglicht breiteren / schmaleren Suchfokus
  - Klassifikation von Dokumenten

Using WordNet in Text retrieval

> Text-Suche als Vektorraummodell

Mehrdeutigkeit bei der Suche:

> **Homographen** verringern Precision

> **Synonyme** verringern Recall

Concept Matching:

> Queries und Dokumente als semantic concordances

> Semantic Tagging mit Hilfe von „Hoods“

> automatisches „Semantic Tagging“ insb. bei Query ein Problem

Query Expansion:

> Synsets aus der Umgebung werden der hinzugefügt.

→beide Verfahren scheitern am „Semantic Tagging“

Meinetz-Suche

Korpus von Meinetz-Suchanfragen. (n=93) .doc

lexikon-tauglich	37
Fächerbezeichnungen	16
nicht erfassbar	62
Eigennamen	42
spezielle Pragmatik	20

Integration in Meinetz

Wie kann WordNet-Wissen in Meinetz verwendet werden?

> Disambiguierung der Schlüssel – Schlüssel werden nicht durch eine Namen sondern einen Wordnet-Sense bestimmt.

> Auch Wörter in der WordNet-Umgebung eines Begriffs können Zugehörigkeit bestimmen. Aber:

» nur wenn disambiguiert (Chaos: Apfelinfos in Computer-Gruppe)

» am besten mit Metrik (Anzahl der Nodes)

References

http://www.cogsci.princeton.edu/~wn/

» Insbesondere „Five Papers on Wordnet“

Fellbaum, C., (1998). WordNet, an electronical lexical database. Cambride, MA: MIT Press.

» Insbesondere: Voorhees E. M.. Using WordNet for Text Retrieval.

Hayes, B., (1999). The web of words. In American Scientist Volume 87 Number 2, pages 108-112.