

Proseminar Linguistische Annotation

Ines Rehbein und Josef Ruppenhofer

SS 2010



Was ist WordNet?

- WordNet enthält Nomen, Verben, Adjektive und Adverbien, die nach kognitiven Aspekten gruppiert wurden
- Jede Gruppe (Synset) drückt ein neues Konzept aus (Synsets \Rightarrow Lesarten)
- Synsets sind durch konzeptuell-semantische und lexikalische Relationen miteinander verbunden
- Resultat: ein Netzwerk von Worten/Konzepten, die miteinander durch semantische Relationen verbunden sind
- WordNet ist ein Wörterbuch, das nach psycholinguistischen Gesichtspunkten erstellt wurde

WordNet = Hierarchie von Konzepten + Relationen zu anderen Konzepten

Was ist WordNet?

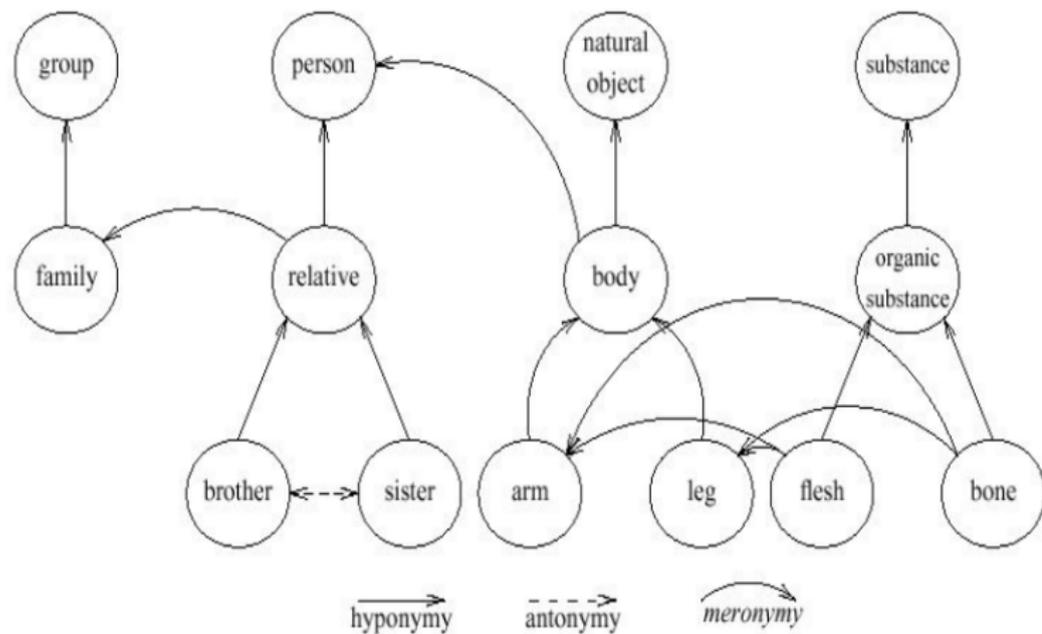
- WordNet enthält Nomen, Verben, Adjektive und Adverbien, die nach kognitiven Aspekten gruppiert wurden
- Jede Gruppe (Synset) drückt ein neues Konzept aus (Synsets \Rightarrow Lesarten)
- Synsets sind durch konzeptuell-semantische und lexikalische Relationen miteinander verbunden
- Resultat: ein Netzwerk von Worten/Konzepten, die miteinander durch semantische Relationen verbunden sind
- WordNet ist ein Wörterbuch, das nach psycholinguistischen Gesichtspunkten erstellt wurde

WordNet = Hierarchie von Konzepten + Relationen zu anderen Konzepten

WordNet - Relationen

- Synonymie (Schiff, Boot)
- Antonymie (heiß, kalt)
- Hyponymie (Rose, Blume)
- Meronymie (Fenster, Haus)
- Entailment (schnarchen, schlafen; nur für Verben)
- Has_member (Familie, Kind)
- Has_stuff (Panzer, Stahl)
- ...

WordNet - Fragment



POS	uniq strings	Synsets	avg. senses
			incl./excl. monosem. words
Noun	117.798	82.115	1.2 / 2.8
Verb	11.529	13.767	2.2 / 3.6
Adjective	21.479	18.156	1.4 / 2.7
Adverb	4.481	3.621	1.2 / 2.5
Totals	155.287	117.659	

- Human IAA on the SENSEVAL-2 data (using WordNet 1.7)
 - ▶ 73% agreement on all words
 - ▶ 85% agreement for nouns and adjectives (which are less polysemous than verbs)

WordNet - Wozu?

- Word Sense Disambiguation (WSD, Lesartenunterscheidung)
 - ▶ verschiedene Lesarten → verschiedene Synsets
- Information Retrieval
 - ▶ Disambiguierung von Fragen/Texten
 - ▶ Erweiterte Suchabfragen mit Hilfe von Synonymen
⇒ besserer Recall
- Klassifikation von Dokumenten oder Konzepten aufgrund von semantischer Ähnlichkeit
 - ▶ Distanz 2er Konzepte definiert durch die Pfadlänge in der WordNet-Hierarchie
 $dist_{WN} = pathlength(s_1, s_2)$
 - ▶ Verschiedene Maße von WordNet-Ähnlichkeit/Distanz (Resnik, Jiang and Conrad, Hirst and St. Onge, Lin, Leacock and Chodorow)
- Machine Translation, Image Retrieval, Audio and Video Retrieval, Language Teaching and Translation, ...

Unterschiede: WordNet - FrameNet

Wordnet	FrameNet
einfache Muster (frames) wie NP verb PP	Korpusbelege für jede Lesart und für grammatikalische Varianten aller Bedeutungen
nach Wortarten getrennt	Beziehungen zwischen Worten des gleichen Frames auch über Wortarten hinaus sichtbar verschiedene POS im gleichen Frame
Wörterbuch + Thesaurus	Wörterbuch + Thesaurus + Verbindungen zwischen Argumentstrukturen von miteinander in Beziehung stehenden Lemmas