

Sémantická analýza WordNet

PLIN059

Mgr. Dana Hlaváčková, Ph.D.

Sémantická analýza

- sémantika (významy slov, slovních spojení)
- sémantické vztahy (hyponymie – hyponymie; kohyponymie)
- často navazuje na analýzu morfologickou a syntaktickou
 - min. určení slovních druhů (entity a abstrakce – substantiva, děje – slovesa, vlastnosti – adjektiva)
- snaha o formální popis významů (jazykově nezávislý)

- ontologie
- pojmenované entity (information retrieval)
- analýza sentimentu
- sémantické sítě

Ontologie

- **ontologie** – významové struktury, skládají se z tzv. konceptů
 - mělká (shallow)
 - strukturovaná, hierarchická
 - **vrcholová** (top ontology, upper ontology)
 - SUMO <http://www.adampease.org/OP/>
 - **doménová** (znalostní obor, terminologie, taxonomie)

Pojmenované entity

- named entity
- hledání předem definovaných kategorií v nestrukturovaném textu
- „sémantické nálepky“ ke slovům, slovním spojením, číslům a znakům
- Named Entity Recognition (NER, CZPJ FI MU)
 - <https://nlp.fi.muni.cz/projekty/ner/v2/>
- Czech Named Entity Corpus (CNEC, ÚFAL MFF UK)
 - <http://ufal.mff.cuni.cz/cnec/cnec2.0>

Sémantické sítě

- FrameNet
 - <https://framenet.icsi.berkeley.edu/fndrupal/>
- VerbNet
 - <https://verbs.colorado.edu/~mpalmer/projects/verbnet.html>
- WordNet
 - <http://wordnet.princeton.edu>

WordNet – motivace

- **G. A. Miller** (*Princeton University*) – psycholog a psycholinguista, psycholexikologie
- uspořádání významů v lidské lexikální paměti
- hierarchie
- experimenty – asociační testy, schopnost zpracovávat anaforické výrazy

WordNet – struktura

- model lexikální paměti, sémantická síť
- **synset** – synonymická řada (blízká synonyma), **literál** + číslo významu
- substantiva, adjektiva, verba, adverbia
- **hierarchická struktura** – hyperonyma, hyponyma, kohyponyma
- substantiva – tematické hierarchie
- verba – vztah vyplývání
- další **sémantické vztahy** – antonyma, holonyma, meronyma, domény ontologií SUMO, MILO
- derivační vztahy

WordNet – projekty

- **Princeton WordNet** (1990–1995) – americká angličtina, G. A. Miller, Christiane Fellbaum
- **EuroWordNet** – Piek Vossen, *University of Amsterdam*
 - **EWN I** (1997–1998) – *angličtina, holandština, italština, španělština*
 - **EWN II** (1998–1999) – *francouzština, němčina, čeština, estonština*
- **Balkanet** (2001–2004), D. Christodoulakis, **University of Patras** – *turečtina, rumunština, řečtina, bulharština, srbština, čeština*

EuroWordNet

- **Base Concepts** – jádro slovní zásoby (cca 1000 synsetů)
- **Top-Ontology** – 63 konceptů
 - entity 1. řádu = objekty
 - entity 2. řádu = stavy a procesy
 - entity 3. řádu = abstraktní pojmy (množina)
- **Interlingual Index** – číslo, které propojuje významy v jednotlivých wordnetech

Odkazy

- Global WordNet Association
 - <http://www.globalwordnet.org>
- nástroje, prohlížeče
 - VisDic, DEBVisDic (doplněk Firefoxu), DEBVisDic 2 (webové rozhraní)
 - Responsivity Aware Wordnet viewer (RAW)
 - <https://deb.fi.muni.cz/raw-viewer/rawviewer.html>