Einführung in die Computerlinguistik und Sprachtechnologie

Vorlesung im WS 2009/10

Prof. Dr. Udo Hahn
Lehrstuhl für Computerlinguistik
Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena

Computational Lexicons, Terminologies & Ontologies

• Computational Lexicons
  – WordNet (English) & EuroWordNet
  – FrameNet
• (Biomedical) Terminologies
  – Unified Medical Language System (UMLS)
  – GENIA Ontology
  – Open Biological Ontologies (OBO)
    • Gene Ontology (GO)
• Ontologies
  – Formal Reasoning

WordNet

• English WordNet (V3.0)
  – semantic (relation) lexicon of English (general language)
    • no morphology!, no syntax!, no etymology
  – groupings of words into sets of synonyms (synsets)
  – English definitions for lexical entries/ synsets (glosses)
  – defines semantic relations between synsets
  – covers (base forms of) nouns, verbs, adjectives, adverbs
  – Size: more than 155,000 lexical entries

http://wordnet.princeton.edu/
**WordNet**

- EuroWordNet
  - Portuguese, Spanish, Spanish Catalan-Basque
  - Russian
  - French
  - German (poor coverage, fees are charged)
  - ...

- Global WordNet [http://globalwordnet.org](http://globalwordnet.org)
  - Arabic
  - Mandarin-Chinese
  - Hindi
  - ...

**WordNet SynSets and Gloses**

**Nouns**

S: (n) jump, leap (a sudden and decisive increase) "a jump in attendance"
  - direct hyponym | full hyponym
S: (n) quantum leap, quantum jump (a sudden large increase or advance) "this may not insure success but it will represent a quantum leap from last summer"
  - direct hyponym | inherited hyponym | sister term (derivationally related form)
S: (n) leap, jump, saltation (an abrupt transition) "a successful leap from college to the major leagues"
  - S: (n) starlin, jump, start (a sudden involuntary movement) "he awoke with a start"
S: (n) jump, parachuting (descent with a parachute) "he had done a lot of parachuting in the army"
  - S: (n) jump, jumping (the act of jumping; propelling yourself off the ground) "he advanced in a series of jumps"; "the jumping was unexpected"

**Verb**

S: (v) jump, leap, bound, spring (move forward by leaps and bounds) "The horse bounded across the meadow"; "The child leapt across the puddle"; "Can you jump over the fence?"
  - S: (v) startle, jump, start (move or jump suddenly, as if in surprise or alarm) "She startled when I walked into the room"
S: (v) jump (make a sudden physical attack on) "The muggers jumped the woman in the fur coat"
  - S: (v) jump (increase suddenly and significantly) "Prices jumped overnight"
S: (v) jump, leap, jump out, jump out, jump, stand out, stick out (be highly noticeable) "the parachutist didn't want to jump"; "every year, hundreds of people jump off the Golden Gate bridge"; "the widow leapt into the funeral pyre"
  - S: (v) derail, jump (run off or leave the rails) "the train derailed because a cow was standing on the tracks"
S: (v) chute, parachute, jump (jump from an airplane and descend with a parachute)
  - S: (v) jump, leap (cause to jump or leap) "the trainer jumped the tiger through the hoop"
S: (v) jumpstart, jump-start, jump (start a car engine whose battery is dead) by connecting it to another car's battery"
  - S: (v) pass over, skip over, skip over (bypass) "He skipped a row in the text and so the sentence was incomprehensible"
S: (v) leap, jump (pass abruptly from one state or topic to another) "leap into fame"; "jump to a conclusion"; "jump from one thing to another"
  - S: (v) alternate, jump (to alternate and not show preference between two sides or viewpoints)
**WordNet Relations**

- **Verbs**
  - **Hypernyms**
    - "the verb Y is a hypernym (more general term) of the verb X, if the activity X is a (kind of) Y"
      - e.g., travel to movement
  - **Troponyms**
    - "the verb Y is a troponym of the verb X, if the activity Y is doing X in some manner"
      - e.g., lisp to talk
  - **Entailment**
    - "the verb Y is entailed by the verb X, if by doing X you must be doing Y"
      - e.g., snoring by sleeping
  - **Coordinate terms**
    - "Y is a coordinate verb of X, if X and Y share a hypernym"

- **Adjectives**
  - Related nouns
  - Participle of verb

- **Adverbs**
  - Root adjectives

**WordNet V3.0 Statistics**

<table>
<thead>
<tr>
<th>POS</th>
<th>Unique Strings</th>
<th>SynSets</th>
<th>Word-Sense Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>117,100</td>
<td>82,100</td>
<td>146,300</td>
</tr>
<tr>
<td>Verb</td>
<td>11,500</td>
<td>13,800</td>
<td>25,000</td>
</tr>
<tr>
<td>Adj</td>
<td>21,500</td>
<td>18,200</td>
<td>30,000</td>
</tr>
<tr>
<td>Adv</td>
<td>4,500</td>
<td>3,600</td>
<td>5,600</td>
</tr>
<tr>
<td>Σ</td>
<td>155,30</td>
<td>117,700</td>
<td>206,900</td>
</tr>
</tbody>
</table>

**WordNet V3.0 Statistics**

<table>
<thead>
<tr>
<th>POS</th>
<th>Monosemous Words / senses</th>
<th>Polysemous Words</th>
<th>Polysemous sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>101,900</td>
<td>16,000</td>
<td>44,400</td>
</tr>
<tr>
<td>Verb</td>
<td>6,300</td>
<td>5,300</td>
<td>18,800</td>
</tr>
<tr>
<td>Adj</td>
<td>16,500</td>
<td>5,000</td>
<td>14,400</td>
</tr>
<tr>
<td>Adv</td>
<td>3,700</td>
<td>700</td>
<td>1,800</td>
</tr>
<tr>
<td>Σ</td>
<td>128,400</td>
<td>27,900</td>
<td>79,500</td>
</tr>
</tbody>
</table>

http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html
WordNet V3.0 Statistics

<table>
<thead>
<tr>
<th>POS</th>
<th>Average polysemy* (incl. monosemous words)</th>
<th>Average polysemy* (excl. monosemous words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>1.24</td>
<td>2.79</td>
</tr>
<tr>
<td>Verb</td>
<td>2.17</td>
<td>3.57</td>
</tr>
<tr>
<td>Adj</td>
<td>1.40</td>
<td>2.71</td>
</tr>
<tr>
<td>Adv</td>
<td>1.25</td>
<td>2.50</td>
</tr>
</tbody>
</table>

* Number of synsets that contain the word

FrameNet

- English FrameNet
  - semantic frames of English (script-style)
    - no morphology!, no syntax!, no etymology
  - English definitions for lexical entries
  - Statistics
    - Version 1.3
    - 11,000 lexical units
    - 970 semantic frames
    - 135,000 example sentences for frames (taken from the British National Corpus [BNC] and US newswire)

[FrameNet: http://framenet.icsi.berkeley.edu/]
[FrameNet: http://framenet.icsi.berkeley.edu/index.php?option=com_content&task=view&id=17881&Itemid=66/]

Try out: FrameGrapher

FrameNet Entry

FrameNet Data Search for jump

Lexical unit search results: Closest match is jump...

<table>
<thead>
<tr>
<th>Lexical Unit</th>
<th>Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>jump.v</td>
<td>Self_motion</td>
</tr>
<tr>
<td>jump.v</td>
<td>Traversing</td>
</tr>
<tr>
<td>jump.v</td>
<td>Change_position_on_a_scale</td>
</tr>
<tr>
<td>jump.v</td>
<td>Attack</td>
</tr>
<tr>
<td>jumper.n</td>
<td>Clothing</td>
</tr>
<tr>
<td>jumping.a</td>
<td>Lively_place</td>
</tr>
<tr>
<td>jumpsuit.n</td>
<td>Clothing</td>
</tr>
</tbody>
</table>

FrameNet Entry (cont.)

Self_motion

Definition: The Self_mover, a living being, moves under its own power in a directed fashion, i.e. along what could be described as a Path, with no separate vehicle.

FES:
Core:
Area [Area] Semantic Type Location
Area is used for expressions which describe a general area in which motion takes place when the motion is understood to be irregular and not to consist of a single linear path. Note that this FE should not be used for cases when the same phrase could be used with the same meaning with a non-motion target, since these should be annotated with the Place FE.
Direction [dir]
The direction that the Self_mover heads in during the motion.
Goal [Goal] Semantic Type Goal
Goal is used for any expression which tells where the Self_mover ends up as a result of the motion.
Path [Path] Semantic Type Path
Path is used for any description of a trajectory of motion which is neither a Source nor a Goal. This includes "middle of path" expressions.
Self_mover [SMov] Semantic Type Sentient
Self_mover is the living being which moves under its own power. Normally it is expressed as an external argument.
Source [Src] Semantic Type Source
Source is used for any expression which implies a definite starting-point of motion. In prepositional phrases, the prepositional object expresses the starting point of motion. With particles, the starting point of motion is understood from context.
**Spanish FrameNet**

- 1,000 lexical items (Dec 2009)
- Based on 350M word corpus (new World (60%) and European Spanish (40%))

http://gemini.uab.es:9080/SFNsite/

**Biomedical) Terminologies**

- Sublanguages: domain-specific
- Relational Encoding
  - Is-a
  - Part-of

**UMLS – Unified Medical Language System**

- Purpose: clinical coding, billing, …
- Umbrella system made up of more than 100 terminologies
- Size: 2,000,000 terms; 900,000 concepts, 12,000,000 relations
- Content: (almost) the whole of (clinical) medicine
- Lexical semantics: thesaurus relations for taxonomies, partonomies, also other light-weight semantics (approximately 80 additional relation types)
- Basic and variant word forms, and (quite complex) NPs
- (English) Specialist Lexicon uses conceptual grounding of UMLS for NLP applications
UMLS Tables

<table>
<thead>
<tr>
<th>Concept 1</th>
<th>relation</th>
<th>Concept 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>RIGHT-SIDE-OF-HEART</td>
<td>narrower_rel</td>
<td>HEART</td>
</tr>
<tr>
<td>LEFT-SIDE-OF-HEART</td>
<td>part_of</td>
<td>HEART</td>
</tr>
<tr>
<td>ANGINA-PECTORIS</td>
<td>has_location</td>
<td>HEART</td>
</tr>
<tr>
<td>HEART</td>
<td>has_part</td>
<td>HEART-ATRIUM</td>
</tr>
<tr>
<td>WALL-OF-HEART</td>
<td>part_of</td>
<td>HEART</td>
</tr>
<tr>
<td>BRONCHIAL-TUBERCULOSIS</td>
<td>has_location</td>
<td>BRONCHI</td>
</tr>
<tr>
<td>BRONCHIAL-TUBERCULOSIS</td>
<td>narrower_rel</td>
<td>TUBERCULOSIS</td>
</tr>
<tr>
<td>SARCOMA</td>
<td>sibling</td>
<td>CARCINOMA</td>
</tr>
<tr>
<td>LENS-CRYSTALLINE</td>
<td>part_of</td>
<td>EYE</td>
</tr>
<tr>
<td>ACUTE-MYELOID-LEUKEMIA</td>
<td>has_location</td>
<td>BONE-MARROW</td>
</tr>
<tr>
<td>RIGHT-HAND</td>
<td>is_a</td>
<td>HAND</td>
</tr>
<tr>
<td>ALLERGIC-REACTION</td>
<td>associated_with</td>
<td>DERMATITIS-ATOPIC</td>
</tr>
<tr>
<td>LUNG</td>
<td>broader_rel</td>
<td>ATELECTASIS</td>
</tr>
</tbody>
</table>

anatomical concepts
pathological concepts

GENIA Ontology

- **Purpose:**
  - biological named entity annotation (prerequisite subroutine for text mining)

- **Coverage:**
  - Cell signaling reactions in human
    - substances involved in biochemical reactions
    - biological locations where substances are found & reactions take place (e.g., organisms, tissues, cells)

- **45 categories**
  - Informal semantics:
    - verbal "scope notes" as an informal phrasing of the categories' meaning (no definitary axioms)

- **1 base relation (is-a), 44 relation instances**
  - [GENIA Ontology](http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/genia-ontology.html)

The Complete GENIA Ontology

GENIA Ontology Scope Notes

"An individual member of a group of non-complex proteins, e.g., STAT1, STAT2, STAT3, or a (non-complex) protein not regarded as a member of a particular group."

NF kappa B, CD28, IL-3, GTP-bound p21ras, HIV-1 tat, protein kinase C, (…)

"A family or a group of proteins, e.g., STATs"

antibodies, transcription factors, cytokines, cytosolic protein, T-cell receptors, DNA binding protein, (…)
Open Biological Ontologies (OBO)

http://obo.sourceforge.net

- Coverage:
  - Anatomy (cells, human, model organisms, etc.)
  - Chemical entities
  - Experimental conditions
  - Genomics, proteomics
  - ...
- Structured controlled vocabularies (thesauri)
- Basic Relations: is-a, part-of
- OBO entry: ID, concept name, textual definition, synonyms

Gene Ontology (GO)

- Purpose: Data annotation and integration for genes and gene products (cross-species)
- Coverage: Three ontologies in one for molecular biology
  - cellular component: location of a gene product, within (sub)cellular structures and macromolecular complexes, e.g., nucleus or ribosome
  - molecular function: the tasks performed by individual gene products at the biochemical level, e.g., enzyme or transporter
  - biological process: biological goals to which a gene product contributes; that process is accomplished by ordered assemblies of molecular functions, e.g., mitosis or cell growth
- 20,500 categories (95.6% w. verbal definitions)
- 2 base relations; 30,500 relation instances
  - Specific/general (88%) (mitotic chromosome is-a chromosome)
  - Part/whole (12%) (telomere part-of chromosome)

OBO Statistics

(Dec 2009)

- More than 60 OBO ontologies
  - about 50% of them contain more than 1000 terms:
    - 2 x > 25 000 terms: NCI Thesaurus, FMA (Human Anatomy)
    - 5 x 10 000-25 000: GO (total), disease ontology, MeSH “ontology”, mouse anatomy stages, ChEBI (chemicals)
    - 18 x 1000 -10 000 terms: molecule role (chemicals, protein by function), human, mouse, fly, fish anatomy (some: developmental anatomy), phenotype ontology, tissue ontology, sequence ontology
- Less than 1000 terms: cell ontology, pathway ontology, MGED (Microarray Gene Expression Database), relationship ontology (amongst others)
- Rapidly growing! – check out every day (o.k., week is also fine)

http://www.obofoundry.org

Snapshot of GO

is-a relation

mitochondrion

Accession: GO:0005739
Ontology: cellular_component
Definition: A semi-autonomous, cell replicating organelle that occurs in varying numbers, shapes, and sizes in the cytoplasm of virtually all eukaryotic cells. It is notably the site of tissue respiration.

part-of relation

mitochondrial membrane

Accession: GO:0005741
Ontology: cellular_component
Definition: A thin double layer of phospholipid bilayer that surrounds all eukaryotic nuclei and organelles.
General Shortcomings

- Category descriptions, at best, are verbally defined
- Relations are usually undefined, their names appeal to human/expert intuition
- (Almost) No attempt at interoperability
- Lots of unlinked fragments (still a long way to go to some sort of ‘Bio-UMLS’)

Ontologies

- Formal Reasoning
- Conceptual Computation

Why Conceptualize?

- Nomenclatures, thesauri, ontologies, …
- “Mapping problem” due to term variation
  - Natural language $\Rightarrow$ domain knowledge

“Mapping Problem” (1/2)

Problem: Mapping a textual occurrence of a bio entity (text token, term) to its ontological category (type)

- Orthographic variations
  - Hyphens, slashes, spaces (e.g., NF-KB, NF KB, NF/KB, NFKB)
  - Upper/lower cases (e.g., NF-KB, NF-kb)
  - Spelling variations (e.g., tomour vs. tumor, oestrogen vs. estrogen, alpha vs. $\alpha$)
- Lexical and phrasal variations
  - Acronyms (e.g., RAR vs. retinoic acid receptor)
  - Different reductions (e.g., SB2 gene vs. SB2, thyroid hormone receptor vs. thyroid receptor)
“Mapping Problem” (2/2)

- Semantic variations (n:m token-type relations)
  - n:1 Synonyms (e.g., in FlyBase: EST-6 vs. Esterase 6 vs. carboxyl ester hydrolase)
  - 1:m Ambiguity as polysemy (e.g., ‘per’ in FlyBase: period gene vs. clock gene)

Why Is Bio Terminology So Hard?

Why Conceptualize?

- Nomenclatures, thesauri, ontologies, ...
- “Mapping problem” due to term variation
  - Natural language vs. domain knowledge
- “Structure computing” on knowledge structures
  - Lexical look-up
  - Relational navigation (general-specific, is-a)
  - Formal reasoning (inferencing)

“Structure Computing”
How Things Got Started …

Tree of life web project:
http://tolweb.org/tree/phylogeny.html
“Structure Computing”
... and Where We Are Heading to

Why Conceptualize?
- Nomenclatures, thesauri, ontologies, ...
- “Mapping problem” due to term variation
  - Natural language \(\rightarrow\) domain knowledge
- “Structure computing” on knowledge structures
  - Lexical look-up
  - Relational navigation (general-specific, is-a)
  - Formal reasoning (inferencing)
- Bio view: data annotation & data integration

Bio View:
Swiss-Prot and GO Terms

Ontologies and Data Annotation

http://www.expasy.org/sprot/
Ontologies and Data Integration

Why Conceptualize?
- Nomenclatures, thesauri, ontologies, ...
- “Mapping problem” due to term variation
  - Natural language ⇔ domain knowledge
- “Structure computing” on knowledge structures
  - Lexical look-up
  - Relational navigation (general-specific, is-a)
  - Formal reasoning (inferencing)
- Bio view: data annotation & data integration
- NLP view: text-based content management
  - Category classification (IR)
  - Semantic interpretation (IE, TM)

NLP view:
Two Text-Based CM Paradigms

Information Extraction

Thalidomide was found to be highly effective in managing the cutaneous manifestations of leprosy (erythema nodosum leprosum) and even to be superior to aspirin (acetylsalicylic acid) in controlling leprosy-associated fever.
S1 A mitochondrion provides the cell with energy in the form of ATP.

S2 The organelle possesses its own genetic material which is inherited maternally.

S3 The ATP synthesizing enzyme ATP synthase is located in the inner membrane.
Conceptual Normalization

S1 A mitochondrion provides the cell with energy in the form of ATP.

S2 The mitochondrion possesses its own genetic material which is inherited maternally.

S3 The ATP synthesizing enzyme ATP synthase is located in the mitochondrial inner membrane.

Semantic Interpretation

“Normalized” Text Level

S1 A mitochondrion provides the cell with energy in the form of ATP.

Propositional Level

• Provide [mitoch., cell, energy]

S2 The mitochondrion possesses its own genetic material which is inherited maternally.

• Possess [mitoch., gen. material]

S3 The ATP synthesizing enzyme ATP synthase is located in the mitochondrial inner membrane.

• Synthesize [ATP synthase, ATP]
• Located-in [ATP synthase, mitoch. inner membrane]

Reasoning on Medical Ontologies

1. Taxonomy

"is-a"

Finger

is-a

Thumb

is-a

Left Thumb

Reasoning on Medical Ontologies

2. Mereology

"part-of"

Thumb

part-of

Thumbnail
Reasoning on Medical Ontologies

1. Taxonomy
2. Mereology
   "part-of"

Reasoning on Bio Ontologies

1. Taxonomy
   "is-a"

2. Mereology
   "part-of"
Reasoning on Bio Ontologies

1. Taxonomy
   "is-a"

2. Mereology
   "part-of"

---

Cell Cycle

Metaphase Mitosis

Mitosis

Cell Cycle

Metaphase

---

Recommended Readings

- Computational Lexicons

- Ontologies
  - S. Staab & R. Studer (Eds.) (2004), Handbook on Ontologies. Springer Publisher