Ontologizing Senses

2009-02-24

Ed Hovy
Idea and Approach

Word1
- Sense1,1
- Sense1,2
- Sense1,3

Word2
- Sense2,1
- Sense2,2
- Sense2,3

Word3
- Sense3,1
- Sense3,2

Word4
- Sense4

Word5
- Sense5

Word6
- Sense6

Sense creation
Verified by annotation

Pool1
UM1
UM2
UM3
UM4
UM5
UM6
UM7
UM8
UM9
UM10

Pool creation during sense creation
Verified by agreement

Pool insertion (ontologizing)
Verified by agreement

Pool merging
Verified by agreement
### Example Sense Pools

**Weapon - instrument used for combat**

<table>
<thead>
<tr>
<th>Sense ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>weapon.a.n.1</td>
<td>Instrument for combat</td>
</tr>
<tr>
<td>weapon.o.n.1</td>
<td>Instrument for combat</td>
</tr>
<tr>
<td>weapon.o.n.3</td>
<td>Instrument for combat</td>
</tr>
<tr>
<td>arsenal.o.n.1</td>
<td>Collection of weapons</td>
</tr>
<tr>
<td>arm.a.n.2</td>
<td>Armament, weaponry</td>
</tr>
<tr>
<td>arm.o.n.7</td>
<td>Armament, weaponry</td>
</tr>
</tbody>
</table>

**Barrier - artifact situated to prevent access**

<table>
<thead>
<tr>
<th>Sense ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>barrier.a.n.1</td>
<td>Physical barrier to prevent access</td>
</tr>
<tr>
<td>barrier.o.n.1</td>
<td>Physical barrier to prevent access</td>
</tr>
<tr>
<td>barrier.o.n.2</td>
<td>Physical barrier to prevent access</td>
</tr>
<tr>
<td>obstruction.a.n.1</td>
<td>Obstruction to prevent access</td>
</tr>
<tr>
<td>obstruction.o.n.3</td>
<td>Obstruction to prevent access</td>
</tr>
<tr>
<td>impediment.o.n.4</td>
<td>Impediment to prevent access</td>
</tr>
</tbody>
</table>

**Commentary**

- An instrument of fighting or combat.
- A physical artifact that impedes or restricts free movement, or denies access to some area.
Possible Uses of Ontology-Based Information

- Pool contains multiple words’ senses:
  - e.g., barrier.a.n.1, obstruction.a.n.1, impediment.o.n.3
  - (Quasi-)synonyms extends training data for WSD, etc.

- Pool contains Chinese and Arabic wordsenses:
  - e.g., xin1hua2she4, da4xue2
  - Ditto, plus may help MT at sense-aligned level

- Pools arranged under Upper Model:
  - e.g., different kinds of Artifacts
  - Closely similar yet different pools still share some type characteristics; perhaps useful for metonymy or gross WSD

- Pool contains features:
  - e.g., +PHYSICAL, +CONCRETE, -MOVEMENT
  - Eventually may be used for semantic role labeling/assignment
  - Features will be inherited
Upper Model

- **Objects (for noun pools):** Stable at about 128 concepts
- **Events (for verb pools):** Still growing, now about 20 concepts

**Pool attachment:**
- Each UM concept is potential ‘attachment point’ for sense pools
- Most pools attach at UM fringe: about 70 for Objects and 15 for Events

**Planned eventual structure:**
- Fairly dense clusters of pools under UM attachments, no deep taxonomies
- Many uplinks for event pools, fewer for objects
- Pools within a cluster share some features, and have some unique ones
- Features: +/-Animate, +/-Concrete, etc.
Pool Ontologizing (English)

- Each pool seen by between 3 and 6 people independently
- Weekly or biweekly telecons and online discussion
- 200 hard cases done by experts
- Accepted only UM attachments with majority agreement

<table>
<thead>
<tr>
<th>Noun sense</th>
<th>UM choice 1</th>
<th># people</th>
<th>ITA</th>
<th>UM choice 2</th>
<th># people</th>
<th>ITA</th>
<th>UM choice 3</th>
<th># people</th>
<th>ITA</th>
</tr>
</thead>
<tbody>
<tr>
<td>channel.y.n.7</td>
<td>Artifact</td>
<td>3</td>
<td>100%</td>
<td>NaturalNonLivingObj</td>
<td>3</td>
<td>100%</td>
<td>LocationOf*NaturalNon</td>
<td>3</td>
<td>100%</td>
</tr>
<tr>
<td>club.y.n.7</td>
<td>LocationOf*Artifact</td>
<td>3</td>
<td>100%</td>
<td>CommercialOrganization</td>
<td>3</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>colony.y.n.2</td>
<td>SocialGathering</td>
<td>3</td>
<td>100%</td>
<td>MeasurementQuantityC</td>
<td>3</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>colony.y.n.5</td>
<td>SpatialLocation</td>
<td>3</td>
<td>100%</td>
<td>SocialAbstraction</td>
<td>3</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>colony.y.n.3</td>
<td>SocialGathering</td>
<td>3</td>
<td>100%</td>
<td>MeasurementQuantityC</td>
<td>3</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>column.y.n.7</td>
<td>ShapeAndStructureA</td>
<td>3</td>
<td>100%</td>
<td>GeneralizedLanguageC</td>
<td>3</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>compartment.</td>
<td>Artifact</td>
<td>3</td>
<td>100%</td>
<td>PartOf*Artifact</td>
<td>3</td>
<td>100%</td>
<td>ShapeAndStructureAbc</td>
<td>3</td>
<td>100%</td>
</tr>
<tr>
<td>core.y.n.8</td>
<td>Human</td>
<td>3</td>
<td>100%</td>
<td>NonProfessionRole</td>
<td>3</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>country.y.n.6</td>
<td>LocationOf*NaturalN</td>
<td>3</td>
<td>100%</td>
<td>SpatialLocation</td>
<td>3</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>crash.y.n.5</td>
<td>ThoughtProcess</td>
<td>3</td>
<td>100%</td>
<td>EventAsObject</td>
<td>3</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cubicle.y.n.5</td>
<td>Artifact</td>
<td>3</td>
<td>100%</td>
<td>ShapeAndStructureAbc</td>
<td>3</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deal.y.n.8</td>
<td>PsychoSocialAbstract</td>
<td>3</td>
<td>100%</td>
<td>EventAsObject</td>
<td>3</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deal.y.n.5</td>
<td>Artifact</td>
<td>3</td>
<td>100%</td>
<td>SetOf*Artifact</td>
<td>3</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>deal.a.n.7</td>
<td>PsychoSocialAbstract</td>
<td>3</td>
<td>100%</td>
<td>QualityAsObject</td>
<td>3</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>debate.y.n.1</td>
<td>EventAsObject</td>
<td>3</td>
<td>100%</td>
<td>GeneralizedLanguageC</td>
<td>3</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>negotiation.y.r</td>
<td>EventAsObject</td>
<td>3</td>
<td>100%</td>
<td>GeneralizedLanguageC</td>
<td>3</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>decline.y.n.5</td>
<td>Artifact</td>
<td>3</td>
<td>100%</td>
<td>NaturalNonLivingObj</td>
<td>3</td>
<td>100%</td>
<td>ShapeAndStructureAbc</td>
<td>3</td>
<td>100%</td>
</tr>
<tr>
<td>department.y.</td>
<td>PoliticoGovernmentC</td>
<td>3</td>
<td>100%</td>
<td>EducationalOrganization</td>
<td>3</td>
<td>100%</td>
<td>MilitaryOrganization</td>
<td>3</td>
<td>100%</td>
</tr>
<tr>
<td>department.y.</td>
<td>PoliticoGovernmentC</td>
<td>3</td>
<td>100%</td>
<td>EducationalOrganization</td>
<td>3</td>
<td>100%</td>
<td>CommercialOrganization</td>
<td>3</td>
<td>100%</td>
</tr>
<tr>
<td>design.y.n.6</td>
<td>Artifact</td>
<td>3</td>
<td>100%</td>
<td>ShapeAndStructureAbc</td>
<td>3</td>
<td>100%</td>
<td>GeneralizedLanguageC</td>
<td>3</td>
<td>100%</td>
</tr>
<tr>
<td>door.y.n.4</td>
<td>PartOf*Artifact</td>
<td>3</td>
<td>100%</td>
<td>LocationOf*Artifact</td>
<td>3</td>
<td>100%</td>
<td>SpatialLocation</td>
<td>3</td>
<td>100%</td>
</tr>
<tr>
<td>plunge.y.n.2</td>
<td>ShapeAndStructureA</td>
<td>3</td>
<td>100%</td>
<td>EventAsObject</td>
<td>3</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>drop.y.n.5</td>
<td>NaturalNonLivingObj</td>
<td>3</td>
<td>100%</td>
<td>ShapeAndStructureAbc</td>
<td>3</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Merging Process

• Task: Merge incoming pool (derived from English monosemous or Arabic or Chinese noun) into appropriate pool, or create new pool

• Approach:
  – Reduce candidates by automated initial assignment to likely pools (heuristics: definition overlaps, WordNet links, etc.)
  – Interface: CAT display tool too cumbersome
  – Developed dedicated interface
  – Speed: fast, over 1 per minute

• Results:
  – English monosemous: 4 annotators, working on initial 2800 terms (covers current corpus); another 7500 in reserve to be done later
  – Chinese: 2 annotators, completed (but not checked) about 300 Chinese pools
  – Arabic: 2 annotators, work underway
Pool Insertion/Merging Interface

Candidate pools for insertion or merging

Details of each candidate pool

Pool to be inserted or merged

Insertion or merging operations
Current Status 1

• Upper Model
  – Objects (from nouns): Stable at about 128 concepts
  – Events (from verbs): Some changes, now about 25 concepts

• Ontologizing English sense pools under Upper Model
  – Pools from polysemous nouns:
    • About 2100 pools linked so far
    • Pools cover about 5300 senses, which (partly) cover about 970 nouns
    • Each pool derived from polysemous English nouns
    • 200 hard cases done by experts; all other linking decisions done by 3 or more people
  – Pools from polysemous verbs (at Columbia):
    • Over 670 pools linked so far (over 700 created)
Current Status 2

- Pools from English monosemous nouns:
  - WordNet senses to UM attachments: Unambig links: 26.2K; ambig links: 895; no link: 10.8K
  - Manual linking/merging of 2800 corpus mono nouns underway

- Chinese noun pools:
  - Pool creation and ontologizing: About 300 pools completed
  - Merging with English-derived pools:
    • Approx. 1200 links created
    • 87 links with perfect agreement, 1137 links with no agreement
  - Chinese verb pools: about 100

- Arabic noun pools:
  - Pool creation and ontologizing: About 320 pools completed
  - Merging with English-derived pools:
    • Approx. 200 links created
Longer-Range Plans

- Continue ontologizing and merging noun- and verb-derived pools for English, Chinese, Arabic
- Finalize taxonomization of pools under each UM node
- Regularize features within pools:
  - Currently 5000 features, about 2000 high-freq
- Finalize links from pools back to WordNet
- Add information from other resources into UM (funded by DoD):
  - Domain labels from NSA
  - Axioms from SUMO
  - Modalities
BACKUP
<table>
<thead>
<tr>
<th>Metric</th>
<th>English</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total noun types double-annotated</td>
<td>1311</td>
<td>200</td>
</tr>
<tr>
<td>Total noun types adjudicated</td>
<td>623</td>
<td>64</td>
</tr>
<tr>
<td>Total polysemous noun instances</td>
<td>141890</td>
<td>52284</td>
</tr>
<tr>
<td>Total noun instances double-annotated</td>
<td>81024 (57.10%)</td>
<td>9557 (18.28%)</td>
</tr>
<tr>
<td>Total noun instances adjudicated</td>
<td>44361</td>
<td>5514</td>
</tr>
<tr>
<td>Average agreement</td>
<td>0.93</td>
<td>0.95</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Arabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>98</td>
</tr>
<tr>
<td>61</td>
</tr>
<tr>
<td>141890</td>
</tr>
<tr>
<td>10201 (7.19%)</td>
</tr>
<tr>
<td>6411</td>
</tr>
<tr>
<td>0.89</td>
</tr>
</tbody>
</table>
Upper Model for Objects

TangibleObject
  Animal
  NonVolitionalBiologicalObject
  VolitionalNonBiologicalObject
  NonVolitionalNonBiologicalObject

Animals: 7 leaf nodes
Other living things: 5 leaf nodes
Androids: 1 leaf node
Artifacts and natural nonliving things: 6 leaf nodes

IntangibleObject
  MentalObject
  Collection
  Relation
  ImmeasurableObject
  MeasurableAbstraction

Thoughts, emotions, etc.: 7 leaf nodes
Sets, groups, organizations: 11 leaf nodes
Relations between things: 11 leaf nodes
Abstractions (information, procedures, functions, etc.): 12 leaf nodes
Space, time, wealth, etc.: 12 leaf nodes
Upper Model for Events  
(Work done at Colorado)

Eventuality1
  Change
    ChangeIntegralProperties
    ChangeIncidentalProperties
Motion
MentalEvent
PhysicalExperience
SocialInteraction

Eventuality2
  Aspect
  Modality

- Change of core defining properties
- Change of other properties: location, possession, etc.
- Motions
- Thoughts
- Bodily experiences
- Communication, interaction, judgment, etc.

- Aspect: completedness, continuity, etc.
- Modality: desires, wishes, obligations, etc.
Pool Ongtologizing Interface (CAT)

- Choices from UM (defs at bottom)
- Item to annotate (def, features) with link to details