Semantic Annotation, Ontology Building, and Interactive Key Generation from Morphological Descriptions

Hong Cui, Alex Dusenbery, James Macklin, Fengqiong Huang, Robert (Bob) Morris, Heather Cole

University of Arizona, UMass-Boston, Agriculture Agri-Food, Canada, University of Arizona, UMass-Boston, Agriculture Agri-Food, Canada

TDWG Beijing, Oct 22- 26 2012
Agenda

- Fine-Grained Semantic Markup
  - CharaParser, a semantic parser for phenotypes
    - Format conversion utilities: XML to RDF and SDD
  - OTO, a web-based application for ontology building
  - CFG, configurable field guide
- Challenges
CharaParser for Fine-Grained Semantic Annotation

- Annotate **factual** information from **textual morphological descriptions** of biodiversity in such a **detailed** manner that the machine readable annotation conveys the information of the original text.

- **Input**: plain text descriptions
- **Output**: XML files with organs/parts, characters, character states, relations, modifiers, and constraints explicitly annotated.
- Employs unsupervised machine learning and syntactic parsing techniques.
“Roots yellow to medium brown or black, thin.”
- root color range = “yellow to medium brown”
- root color = “black”
- root extent = “thin”

“Stems erect to prostrate, often with swollen nodes”
- stem orientation range = “erect to prostrate”
- node solid_shape = “swollen”
- Stems has_part nodes
Machine-Readable Format

- Native format: XML (eXtensible Markup Language)
  - [XML schema for annotation](Web link)
- Example annotation in XML format
- XML annotation can be transformed into other standard formats, such as
  - SDD (Structured Descriptive Data)
  - RDF (Resource Description Framework)
  - Taxon-character matrices
CharaParser Performance

- Performance evaluated on
  - Flora of North America (FNA) volume 19
  - Treatise on Invertebrate Paleontology (TIP) part H.
  - Using precision (P) and recall (R)

- Results published in JASIST 2012.

<table>
<thead>
<tr>
<th></th>
<th>N (complex)</th>
<th>Structure P/R</th>
<th>Character P/R</th>
<th>Relation P/R</th>
<th>Sentence P/R</th>
<th>Correct N</th>
</tr>
</thead>
<tbody>
<tr>
<td>FNA.v19</td>
<td>559 (22%)</td>
<td>99/95</td>
<td>91/90</td>
<td>85/49</td>
<td>96/94</td>
<td>434 (78%)</td>
</tr>
<tr>
<td>TIP.h</td>
<td>457 (42%)</td>
<td>97/97</td>
<td>80/87</td>
<td>79/59</td>
<td>90/90</td>
<td>262 (57%)</td>
</tr>
</tbody>
</table>
From Semantic Annotation to Taxon-Character Matrices

- The XML schema for annotation is based on entity-relation model
  - Entity, attributes, and relations among entities
- Translating XML annotation to taxon-character matrices is straightforward, but
  - Sparse matrix problem
    - Inherit characters/character states from higher taxa
  - Polymorphism
  - Numerical expressions
  - Parenthetical characters
Matrix Generated for ACHillea (Plant Genus)

<table>
<thead>
<tr>
<th></th>
<th>apex_texture</th>
<th>array_architecture</th>
<th>blade_length_from</th>
<th>blade_length_to</th>
<th>blade_shape_from</th>
<th>blade_shape_to</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACHillea</td>
<td>scarios</td>
<td>compact</td>
<td>open</td>
<td>simple</td>
<td>compound</td>
<td>corymbiform</td>
</tr>
<tr>
<td>alpina</td>
<td>scarios</td>
<td>crowded</td>
<td>simple</td>
<td>compound</td>
<td>corymbiform</td>
<td></td>
</tr>
<tr>
<td>millefolium</td>
<td>scarios</td>
<td>simple</td>
<td>compound</td>
<td>corymbiform</td>
<td></td>
<td>0.035</td>
</tr>
<tr>
<td>nobilis</td>
<td>scarios</td>
<td>simple</td>
<td>compound</td>
<td>corymbiform</td>
<td></td>
<td>0.015</td>
</tr>
<tr>
<td>ptarmica</td>
<td>scarios</td>
<td>simple</td>
<td>compound</td>
<td>corymbiform</td>
<td></td>
<td>0.003</td>
</tr>
</tbody>
</table>
Configurable Field Guide

- Created by Robert Morris and students.
- **Input**: taxon-character matrix
- **Output**: interactive field guide
- Ranks characters in their ability to evenly partition the taxa in question.
  - At each partition, maximize information gain.
Species under consideration:

- alpina
- millefolium
- nobilis
- ptarmica

Characters:

- blade_shape_from
- blade_length_from
- palea_size_from
- blade_shape_to
- cypsela_size_from
- blade_width_from
- blade_length_to
- stem_pubescence_from
- blade_width_to

The characters with value linear-lanceolate for blade_shape_from are:

- alpina

The characters with value ovate for blade_shape_from are:

- nobilis

The characters with value linear for blade_shape_from are:

- ptarmica

The characters with value oblong for blade_shape_from are:

- millefolium

Select state:

- linear-lanceolate
- ovate
- linear
- oblong

Submit

You've either found a single species, or there is no character left to distinguish the remaining species.

Home
OTO: Ontology Term Organizer

- [http://biosemantics.arizona.edu/ONTNEW/](http://biosemantics.arizona.edu/ONTNEW/)
  - username: OTOdemo
  - password: OTOdemopass
  - Select demo dataset: OTO_demo.

- Web-based application
  - **Input**: terms/expressions (extracted by CharaParser)
  - **Output**: clean glossaries or raw ontologies.
  - Used by biologists
  - Sort is_a, part_of and order relationships of terms extracted by CharaParser
  - Resolve conflicts based on consensus
## OTO: Group Terms

### Terms:
- netted
- compact
- congested
- lax
- sparser
- rust
- constricted
- acrass
- contractile
- deconstrict
- xerophyl
- false
- funicula
- integumentary
- aplanate
- involucral
- capricho

### Categories:
- arrangement
- coloration
- condition
- count
- course
- dehiscence
- density
- depth
- derivation
- development
- duration
- external texture
- exudation
- fixation
- fragility
- fusion
- germination
- habit
- height
- internal texture
- length
- life stage
- life style
- location
- maturation
- nutrition
- odor
- orientation
- origin
- pattern
- position
- prominence
- reflectance
- relief
- reproduction
- shape
- size
- structure
- taste
- variability
- venation
- width
- volume

### Locations
<table>
<thead>
<tr>
<th>Source (of aggregated)</th>
<th>Detail Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-bit</td>
<td>Heads usually heterogenous (usually radiate) [homogeneous, deciduous, borne singly (on scapiform pseuduncle) [in corolliform, racemiform, or umbelliform arrays, sometimes aggregated in second-order heads].</td>
</tr>
<tr>
<td>228-bit</td>
<td>borne singly (sometimes on scapiform stene) or in corolliform, paniculiform, or racemiform arrays aggregated in second-order heads; florets 1-3 per individual head in Hecactoloba.</td>
</tr>
</tbody>
</table>
**Pubescence:** papillate, minute, glabrous, hairy, bald, balding, barbate, bearded, bristly

**Pubescence-Density Order:** glabrous

**Shape:** cylindric, ovoid, hemispheric, flat, convex, conic, columnar, ovate, lanceolate, linear

**Shape Order:** flat

**Orientation:** erect, prostrate, ascending, spreading, reflexed, appressed, deflexed

**Orientation Order:** appressed, ascending, erect

<table>
<thead>
<tr>
<th>Context</th>
<th>Glossary</th>
<th>Detail Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source (of papillate)</td>
<td>detail sentence</td>
<td>each style usually ringed at base by a nectary; distally 2-branched with stigmatic papillae borne on adaxial face of each branch in 2 separate or contiguous lines or in a continuous band (styles usually not branched in functionally staminate flowers); style branches apically truncate or appended beyond the stigmatic bands or lines; appendages usually papillate to minute</td>
</tr>
</tbody>
</table>
OTO: Admin Tools

<table>
<thead>
<tr>
<th>#</th>
<th>Term</th>
<th>Accepted Decisions</th>
<th>Other Decisions</th>
<th>Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>alkali</td>
<td>variability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>caudal</td>
<td>structure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>cochlear</td>
<td>arrangement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>contorted</td>
<td>shape</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>contortulate</td>
<td>arrangement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>convolute</td>
<td>arrangement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>corrugate</td>
<td>arrangement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>crumpled</td>
<td>arrangement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>equaling</td>
<td>variability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>few</td>
<td>count</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>fever</td>
<td>count</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>hypogastic</td>
<td>development</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>intimate</td>
<td>arrangement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>intimated</td>
<td>arrangement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>intidicate</td>
<td>arrangement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>modified</td>
<td>variability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>formal</td>
<td>variability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>overlapping</td>
<td>arrangement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>plate</td>
<td>structure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>prominent</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary

OTC: Ontology Term Organizer

CharaParser: computable data

Format Converters

CFG: Configurable Field Guide

SDD: Structured Descriptive Data

Taxon-Character Matrix

RDF: Resource Description Framework

Ontologies

Linked Data
Challenges

- Adapting CharaParser from semi-structured sublanguage to natural language and mixed style input
- Building extension ontologies for target domains
- Determining the appropriate level of expressiveness in annotation: how fine-grained is fine enough
- Measuring character-based semantic similarity among taxa
Acknowledgements

- NSF award no. EF-0849982
- Flora of North America Project

- Contact info:
  - hongcui@email.arizona.edu
More Annotation Examples

1. Example, Num-Example (fna, complex)
2. Constraint-Example (Treatise.h, complex)
3. Num-Example (BHL-OCR, simple)
4. Example (Ant-OCR, complex)