Lexical and Statistical Approaches to Acquiring Ontological Relations

Formal Methods for Casual Ontology?

Olivier Bodenreider
Lister Hill National Center for Biomedical Communications
Bethesda, Maryland - USA
Introduction

◆ Biomedical ontologies
  ● Precisely defined (e.g., formal ontology)
  ● Limited size
  ● Built manually

◆ Large amounts of knowledge
  ● Not represented explicitly by symbolic relations
  ● But expressed implicitly
    ▪ By lexico-syntactic relations (i.e., embedded in terms)
    ▪ By statistical relations (e.g., co-occurrence)
  ● Can be extracted automatically
**Formal vs. casual ontology**

- **Formal ontology**
  - Provides a framework for building sound ontologies
  - Too labor-intensive for building large ontologies

- **Casual ontology**
  - Usually unsuitable for reasoning
  - Tools for automatic acquisition available
General framework

- **Ontology learning**
  - [Maedche & Staab, Velardi]
  - ECAI, IJCAI

- **Term variation**
  - [Jacquemin]

- **Terminology / Knowledge**
  - TKE, TIA

- **Knowledge acquisition/capture**
  - K-CAP

- **Information extraction**
Sources of knowledge for casual ontology

- **Long tradition of terminology building**
  - Over 100 terminologies available in electronic format

- **Large corpora available (e.g., MEDLINE)**
  - Entity recognition tools available
    - E.g., MetaMap (UMLS-based)
    - Several for gene/protein names
  - Information extraction methods

- **Large annotation databases available**
  - MEDLINE citations indexed with MeSH
  - Model organism databases annotated with GO
Formal methods for casual ontology

- Lexico-syntactic methods
  - Lexico-syntactic patterns
  - Nominal modification
  - Prepositional phrases
  - Reified relations
  - Semantic interpretation

- Statistical methods
  - Clustering
  - Statistical analysis of co-occurrence data
  - Association rule mining
Lexico-syntactic methods
Synonymy

- **Source: terminology**
- **Lexical similarity**
  - Lexical variant generation program (UMLS)
  - *norm*
- **Limitations**
  - Clinical synonymy vs. Synonymy
  - Molecular biology

[McCray & al., SCAMC, 1994]
Normalization

- Remove genitive: Hodgkin’s diseases, NOS
- Remove stop words: Hodgkin diseases, NOS
- Lowercase: Hodgkin diseases,
- Strip punctuation: hodgkin diseases,
- Uninflect: hodgkin diseases
- Sort words: hodgkin disease

Disease hodgkin
Normalization Example

Normalization

Hodgkin Disease
HODGKINS DISEASE
Hodkin's Disease
Disease, Hodgkin's
Hodkin's, disease
HODGKIN'S DISEASE
Hodkin's disease
Hodgkins Disease
Hodkin's disease NOS
Hodgkin's disease, NOS
Disease, Hodgkins
Diseases, Hodgkins
Hodgkins Diseases
Hodgkins disease
hodgkin's disease
Disease, Hodgkin
Taxonomic relations  Lexico-syntactic patterns

- Source: text corpus
- Example of patterns
  - *Lamivudin is a nucleoside analogue with potent antiviral properties.*
  - *The treatment of schizophrenia with old typical antipsychotic drugs such as haloperidol can be problematic.*

[Hearst, COLING, 1992]
[Fiszman & al., AMIA, 2003]
**Taxonomic relations**  **Nominal modification**

- **Source:** text corpus / terminology
- **Example of modifiers**
  - Adjective
    - *Tuberculous Addison’s disease*
    - *Acute hepatitis*
  - Noun (noun-noun compounds)
    - *Prostate cancer*
    - *Carbon monoxide poisoning*

Terminology: constrained environment (increased reliability)

[Jacquemin, ACL, 1999]
[Bodenreider & al., TIA, 2001]
Reified relations

- **Source:** terminology
- **Example:** reification of *part of*

\[
<X, \text{is-a}, \text{Part of } Y> \\
<X, \text{part-of}, Y>
\]

- **Augmented relations from reified *part-of* relations**
  - Reified: \( <\text{Cardiac chamber, is-a, Subdivision of heart}> \)
  - Augmented: \( <\text{Cardiac chamber, part-of, Heart}> \)

[Zhang & al., ISWC/Sem. Int., 2003]
Prepositional attachment

◆ Source: text corpus / terminology
◆ Example: of
  - *Lobe of lung* → *part of Lung*
  - *Bone of femur* → *part of Femur*
◆ Restrictions
  - Validity of preposition-to-relation correspondence may be limited to a subdomain (e.g., anatomy)
  - Not applicable to complex terms
    - *Groove for arch of aorta* → NOT *part of Aorta*

[Zhang & al., ISWC/Sem. Int., 2003]
Semantic interpretation

- **Source:** text corpus / terminology
- **Correspondence between**
  - Linguistic phenomena
  - Semantic relations
- **Semantic constraints provided by ontologies**

[Navigli & al., TKE, 2002]
[Romacker, AIME, 2001]
[Rindflesch & al., JBI, 2003]
Semantic interpretation

Hemofiltration in digoxin overdose

Hemofiltration treats Digoxin overdose

Therapeutic or Preventive Procedure treats Disease or Syndrome

Antibiotic treats Disease or Syndrome
Medical Device treats Injury or Poisoning
Pharmacologic Substance treats Congenital Abnormality
Ther. or Prev. Procedure treats Disease or Syndrome

Semantic interpretation
Compositional features of terms

- **Lexical items**
  
- **Terms within a vocabulary**
  - Clinical vocabularies
  - Gene Ontology

- **Terms across vocabularies**
  - SNOMED / LOINC
  - GO / ChEBI

- **Lexicon / Terms**
  - Semantic lexicon

References:
- [Baud & al., AMIA, 1998]
- [McDonald & al., AMIA, 1999]
- [Ogren & al., PSB, 2004]
- [Mungall, CFG, 2004]
- [Dolin, JAMIA, 1998]
- [Burgun, SMBM, 2005]
- [Johnson, JAMIA, 1999]
- [Verspoor, CFG, 2005]
Statistical methods
Taxonomic relations  Clustering

- Source: text corpus
- Principle: similarity between words reflected in their contexts
  - Co-occurring words (+ frequencies)
  - Hierarchical clustering algorithms
    - Similarity measure (cosine, Kullback Leibler)
- Can be refined using classification techniques (e.g., k nearest neighbors)

[Faure & al., LREC, 1998]
[Maedche & al., HoO, 2004]
Associative relations

- Source: text corpus / annotation databases
- Principle: dependence relations
  - Associations between terms
- Several methods
  - Vector space model
  - Co-occurring terms
  - Association rule mining
- Limitations: no semantics

[Bodenreider & al., PSB, 2005]
# Similarity in the vector space model

<table>
<thead>
<tr>
<th>Genes</th>
<th>GO terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>g₁</td>
<td>t₁</td>
</tr>
<tr>
<td>g₂</td>
<td>t₂</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>gₙ</td>
<td>tₙ</td>
</tr>
</tbody>
</table>

**Annotation database**

<table>
<thead>
<tr>
<th>GO terms</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>g₁</td>
<td>t₁</td>
</tr>
<tr>
<td>g₂</td>
<td>t₂</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>gₙ</td>
<td>tₙ</td>
</tr>
</tbody>
</table>
Similarity in the vector space model

<table>
<thead>
<tr>
<th>Genes</th>
<th>GO terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t_1 )</td>
<td>( g_1 )</td>
</tr>
<tr>
<td>( t_2 )</td>
<td>( g_2 )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( t_n )</td>
<td>( g_n )</td>
</tr>
</tbody>
</table>

\[ \text{Sim}(t_i, t_j) = t_i \cdot t_j \]
Analysis of co-occurring GO terms

GO terms

<table>
<thead>
<tr>
<th>g1</th>
<th>t1</th>
<th>t2</th>
<th>...</th>
<th>tn</th>
</tr>
</thead>
<tbody>
<tr>
<td>g2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gn</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Annotation database

| t2-t7 | 1 |
| t2-t9 | 1 |
| t7-t9 | 2 |
| ...   |   |

| t5   | 1 |
| t7   | 2 |
| t9   | 2 |
| ...  |   |
Analysis of co-occurring GO terms

Statistical analysis: test independence
- Likelihood ratio test ($G^2$)
- Chi-square test (Pearson's $\chi^2$)

Example from GOA (22,720 annotations)

- **GO:0006955 [BP]** Freq. = 588
- **GO:0008009 [MF]** Freq. = 53

\[
\text{Co-oc.} = 46
\]

<table>
<thead>
<tr>
<th>GO:0006955</th>
<th>present</th>
<th>absent</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>present</td>
<td>46</td>
<td>542</td>
<td>588</td>
</tr>
<tr>
<td>absent</td>
<td>7</td>
<td>21,583</td>
<td>22,132</td>
</tr>
<tr>
<td>total</td>
<td>53</td>
<td>22,125</td>
<td>22,720</td>
</tr>
</tbody>
</table>

**GO:0008009 immune response**

$G^2 = 298.7$

$p < 0.000$
Association rule mining

GO terms

Annotation database

apriori

• Rules: $t_1 \Rightarrow t_2$
• Confidence: $> .9$
• Support: .05
Example of associations (GO)

◆ Vector space model
  ● MF: *ice binding*
  ● BP: *response to freezing*

◆ Co-occurring terms
  ● MF: *chromatin binding*
  ● CC: *nuclear chromatin*

◆ Association rule mining
  ● MF: *carboxypeptidase A activity*
  ● BP: *peptolysis and peptidolysis*
Discussion and Conclusions
Combine methods

◆ Affordable relations
  - Computer-intensive, not labor-intensive

◆ Methods must be combined
  - Cross-validation
  - Redundancy as a surrogate for reliability
  - Relations identified specifically by one approach
    - False positives
    - Specific strength of a particular method

◆ Requires (some) manual curation
  - Biologists must be involved
Limited overlap among approaches

- Lexical vs. non-lexical
- Among non-lexical

[Venn diagram illustrating the overlap between different approaches (VSM, COC, ARM) with numbers indicating the count of overlap in each category.]

[Bodenreider & al., PSB, 2005]
**Reusing thesauri**

- **First approximation for taxonomic relations**
  - No need for creating taxonomies from scratch in biomedicine

- **Beware of purpose-dependent relations**
  - *Addison’s disease* is *Autoimmune disorder*

- **Relations used to create hierarchies vs. hierarchical relations**

- **Requires (some) manual curation**

  [Wroe & al., PSB, 2003]
  [Hahn & al., PSB, 2003]
Formal vs. Casual

◆ Formal ontology
  • Provides a framework for building sound ontologies
  • Too labor-intensive for building large ontologies

◆ Casual ontology
  • Usually unsuitable for reasoning
  • Tools for automatic acquisition available

What is not useful
  • Formal ontology = righteous
  • Casual ontology = sloppy
Formal and Casual

◆ Formal ontology
  ● Provides a framework which can be used as a reference
  ● Help us think clearly (?) about
    ▪ Concepts
    ▪ Relations (e.g., isa: is a kind of / is an instance of)

◆ Casual ontology
  ● Supported by “cheap” (but formal) methods
  ● Extracted from large amounts of data
  ● Helps populating the framework from formal ontology
Combining *formal* and *casual*

**Formal ontology**
- Provides a framework for building sound ontologies
- Too labor-intensive for building large ontologies
- Can benefit from loosely defined ontologies

**Casual ontology**
- Usually unsuitable for reasoning
- Tools for automatic acquisition available
- Can benefit from formal ontology
  - Organization
  - Validation
Casual ontology as a bridge

◆ Casual ontology
  - Speaks the language of biologists
    - Extracted from text or terminologies
  - Passes (part of) the rigorous framework of formal ontology on to biologists

◆ Casual ontologist
  - Not a sloppy ontologist
    - Uses the formal methods of casual ontology
  - Mediator between formal ontology and biology
Medical Ontology Research

Contact: olivier@nlm.nih.gov
Web: mor.nlm.nih.gov

Olivier Bodenreider
Lister Hill National Center for Biomedical Communications
Bethesda, Maryland - USA