Acquiring Ontological Relations From Biomedical Resources

February 3, 2006

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
General framework

- Ontology learning
  - [Maedche & Staab, Velardi]
  - ECAI, IJCAI
- Term variation
  - [Jacquemin]
- Terminology / Knowledge
  - TKE, TIA
- Knowledge acquisition/capture
  - K-CAP
- Information extraction
Types of relations

◆ Lexical relations
  ● Synonymy

◆ Ontological relations
  ● Intra-ontological
    ■ Subsumption (*isa*)
    ■ Meronymy (*part of*)
    ■ [Instantiation]
  ● Trans-ontological
    ■ Dependence relations
    ■ Contingent relations
Types of methods

- 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
Types of objectives

- **Validate ontologies**
  - Compare with asserted knowledge

- **Extend ontologies**
  - With terms extracted from a corpus

- **Enrich ontologies**
  - Maintenance
  - Alignment

- **Link ontologies to other ontologies**
Biomedical resources available

- 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
Ontologies vs. thesauri

- First approximation for taxonomic relations
  - No need for creating taxonomies from scratch in biomedicine
- Beware of purpose-dependent relations
  - Addison’s disease is a Autoimmune disorder
- Relations used to create hierarchies vs. hierarchical relations
- Requires some degree of manual curation

[Wroe & al., PSB, 2003]
[Hahn & al., PSB, 2003]
Overview

- Validate
- Extend
- Enrich
- Link

<table>
<thead>
<tr>
<th>Intra-ontological</th>
<th>Lexico-syntactic</th>
<th>Statistical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjectival modif.</td>
<td>LS patterns</td>
<td></td>
</tr>
<tr>
<td>Reified part of</td>
<td>Prep. attachment</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trans-ontological</th>
<th>Semantic interpretation</th>
<th>Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Vector space model</td>
<td>Enrich</td>
</tr>
<tr>
<td></td>
<td>Co-occurrence anal.</td>
<td>Link</td>
</tr>
<tr>
<td></td>
<td>Assoc. rule mining</td>
<td></td>
</tr>
</tbody>
</table>
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

result: disease hodgkin
Normalization Example

Hodgkin Disease
HODGKINS DISEASE
Hodkin's Disease
Disease, Hodgkin's
Hodgkin's, disease
HODGKIN'S DISEASE
Hodgkin's disease
Hodgkins Disease
Hodgkin's disease NOS
Hodgkin's disease, NOS
Disease, Hodgkins
Diseases, Hodgkins
Hodgkins Diseases
Hodgkins disease
hodgkin's disease
Disease, Hodgkin

normalize

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

- Syntactic analysis
- Mapping to UMLS
- Semantic rules
- Semantic network relationships
- Select matching rule

- 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
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]
Applications  Ontology validation

- 28,851 pairs of terms
  - Original SNOMED term
  - Transformed term (found in UMLS)
- Corresponding relationship in the Metathesaurus
  - Hierarchical in 50% of the cases
  - “Sibling” in 25% of the cases
  - Missing in 25% of the cases

[Bodenreider & al., TIA, 2001]
Lack of structure within a source

diseases of the skin and subcutaneous tissues

cyzema

acute eczema

infantile eczema

acute infantile eczema
Plesionymy

- posttransfusion hepatitis
- posttransfusion viral hepatitis
Applications  To extend ontologies

- 3 M “simple” MEDLINE NPs
- 21,000 already in the Metathesaurus (eliminated)
- 1.3 M (adj+, noun*, head) NPs
- 1.6 M demodified terms
- 125,464 candidate terms

- Manual review
- Relevance of the association: 83%

Applications  To link ontologies

- 2,700 ChEBI entities (27%) identified in some GO term
- 9,431 GO terms (55%) include some ChEBI entity in their names

[Burgun & al., SMBM 2005]
Examples of links ChEBI-GO

◆ iron [CHEBI:18248]
  
  BP  iron ion transport [GO:0006826]
  MF  iron superoxide dismutase activity [GO:0008382]
  CC  vanadium-iron nitrogenase complex [GO:0016613]

◆ uronic acid [CHEBI:27252]
  
  BP  uronic acid metabolism [GO:0006063]
  MF  uronic acid transporter activity [GO:0015133]

◆ carbon [CHEBI:27594]
  
  BP  response to carbon dioxide [GO:0010037]
  MF  carbon-carbon lyase activity [GO:0016830]
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]
1. **Similarity in the vector space model**

### GO terms

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

### Genes

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

**Annotation database**

*Lister Hill National Center for Biomedical Communications*
1. Similarity in the vector space model

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

GO terms

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

Matrix:

<table>
<thead>
<tr>
<th>t₂-t₇</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>t₂-t₉</td>
<td>1</td>
</tr>
<tr>
<td>t₇-t₉</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Diagonal:

| 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

\[
\begin{array}{c|c|c|c}
& \text{present} & \text{absent} & \text{Total} \\
\hline
\text{present} & 46 & 542 & 588 \\
\text{absent} & 7 & 21,583 & 22,132 \\
\text{total} & 53 & 22,125 & 22,720 \\
\end{array}
\]

\[G^2 = 298.7\]
\[p < 0.000\]
Association rule mining

GO terms

<table>
<thead>
<tr>
<th>Genes</th>
<th>t_1</th>
<th>t_2</th>
<th>...</th>
<th>t_n</th>
</tr>
</thead>
<tbody>
<tr>
<td>g_1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g_2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g_n</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Transaction

<table>
<thead>
<tr>
<th>t_2</th>
<th>t_7</th>
<th>t_9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g_2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Annotation database

**Apriori**

- Rules: t_1 => 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
Limited overlap among approaches

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

Venn diagram showing the overlap among different methods (VSM, ARM, COC) with numbers indicating the count of overlaps.
Discussion
Lexico-syntactic vs. statistical

◆ Lexical
  ● Based on terminologies/ontologies
  ● Inferable semantics

◆ Statistical
  ● Based on knowledge bases
  ● No semantics

- Non-redundant, complementary techniques
- Both require some degree of manual curation (semi-automatic techniques)
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
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