A comprehensive data-driven method for ontological term aggregation

Rimma Pivovarov, Noémie Elhadad
Department of Biomedical Informatics, Columbia University, New York, NY

Introduction
Data mining methods often require using ontologies for standardization
The granularity of these ontologies leads to standardized, but sparse datasets
Similar concepts can be aggregated to enrich the sparse data and prevent signal dilution

Methods
Note-Based Similarity

\[ \text{Sim}_{\text{Note}}(\text{C0002622 Amnesia}, \text{C0751295 Memory Loss}) = \text{Cosine Similarity}(V_{\text{C0002622}}, V_{\text{C0751295}}) = .952 \]

Definitional Similarity

\[ \text{Sim}_{\text{Def}}(\text{C0002622 Amnesia}, \text{C0751295 Memory Loss}) = \frac{(18+6)-18}{6} = 1 \]

Ontological Similarity

\[ \text{Sim}_{\text{Onto}}(\text{C0002622 Amnesia}, \text{C0751295 Memory Loss}) = \frac{(9 + 1 + 9)}{3} - 2(.5) = .5333 \]

Results
Search Space Reduction
Only kept pairs with > 75% note-based similarity
Narrowed number of pairs from 24,328,800 to 796

Gold Standard
Annotated by 2 physicians
If at least one annotator says “No” then dissimilar - otherwise, similar
Asked to answer the following question with Yes, Maybe, No
Considering a patient with Chronic Kidney Disease (CKD), from a clinical standpoint, would you say that these two concepts can be used interchangeably?

Discussion
Context Dependence
- “Difficulty Hearing” vs. “Complete Deafness”
  Generally similar
  Given CKD: Difficulty hearing could indicate adverse drug event from a high medication dose for patient with kidney disease

Definitional Similarity (Surprisingly) Good
- This type of measure is commonly used in word sense disambiguation
- Simple and achieves high quality results

Combining All 3 Similarities Produces Best Results
- All three similarity types add unique information
- Higher precision when fully combined than when examined separately

Robust Ontological Method
- Incorporation of all possible relationship types helps us find paths that other methods do not

Literature Cited

Acknowledgements
The authors would like to thank Dr. Herbert Chase, Dr. Michael Owen, and Dr. Hojat Salimzadeh for their similarity annotations and numerous discussions. Rimma Pivovarov is supported by the National Library of Medicine postdoctoral fellowship T15LM010027-19. This study is supported by R01 LM010027 and NLM contract HHSN27620100024C.