Automatically Expanding the Synonym Set of SNOMED CT using Wikipedia

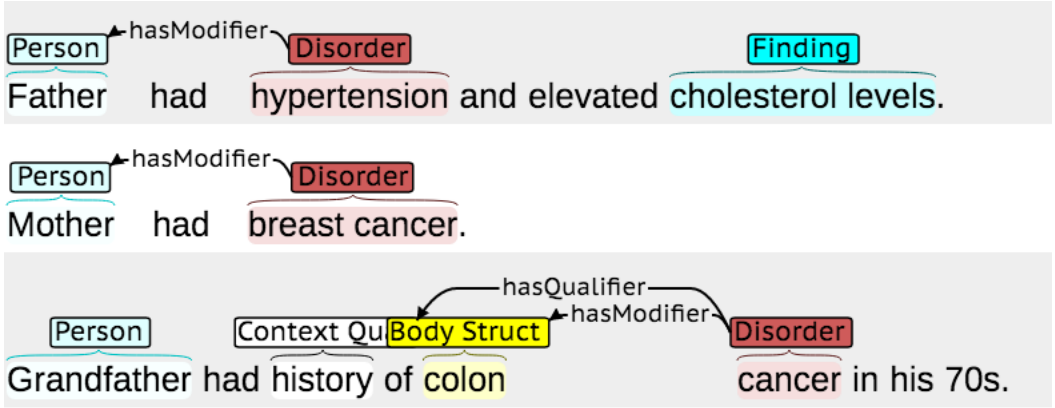
Daniel R. Schlegel, Chris Crowner, and Peter L. Elkin

Department of Biomedical Informatics

University at Buffalo

Our Use Case for SNOMED

- Language Understanding for:
  - Decision support
    - Models for clinical prediction rules, disorders, etc.
  - Information retrieval
    - Inclusion / Exclusion criteria – “5 minute studies”



Research Problem

- Automatic understanding medical text relies on terminologies
  - Coverage is both in primary terms, and synonyms
  - SNOMED CT used very commonly
    - Has about 400,000 terms, 230,000 synonyms
    - -> More synonyms would be helpful
- In some subdomains of medicine, Wikipedia is known to:
  - Have excellent coverage
  - Have similar accuracy to curated web sources
  - Already be used in practice

So, we'd like to leverage Wikipedia to enhance the synonym set of SNOMED CT.

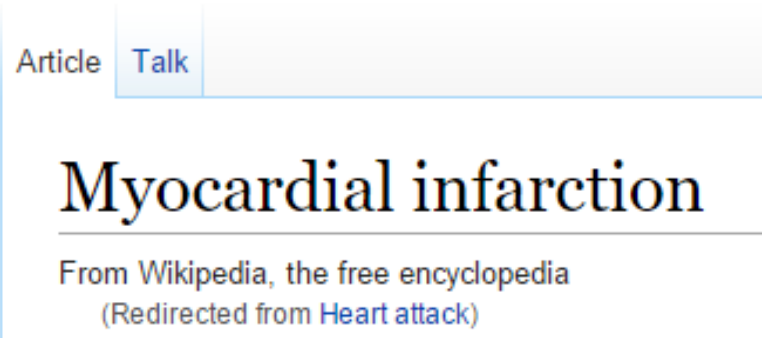
Outline

- Wikipedia Redirects
- Naïve Synonym Harvesting
  - Initial Evaluation
- Problems + Refinements
  - Wikipedia Categories
- Final Evaluation
- Conclusion

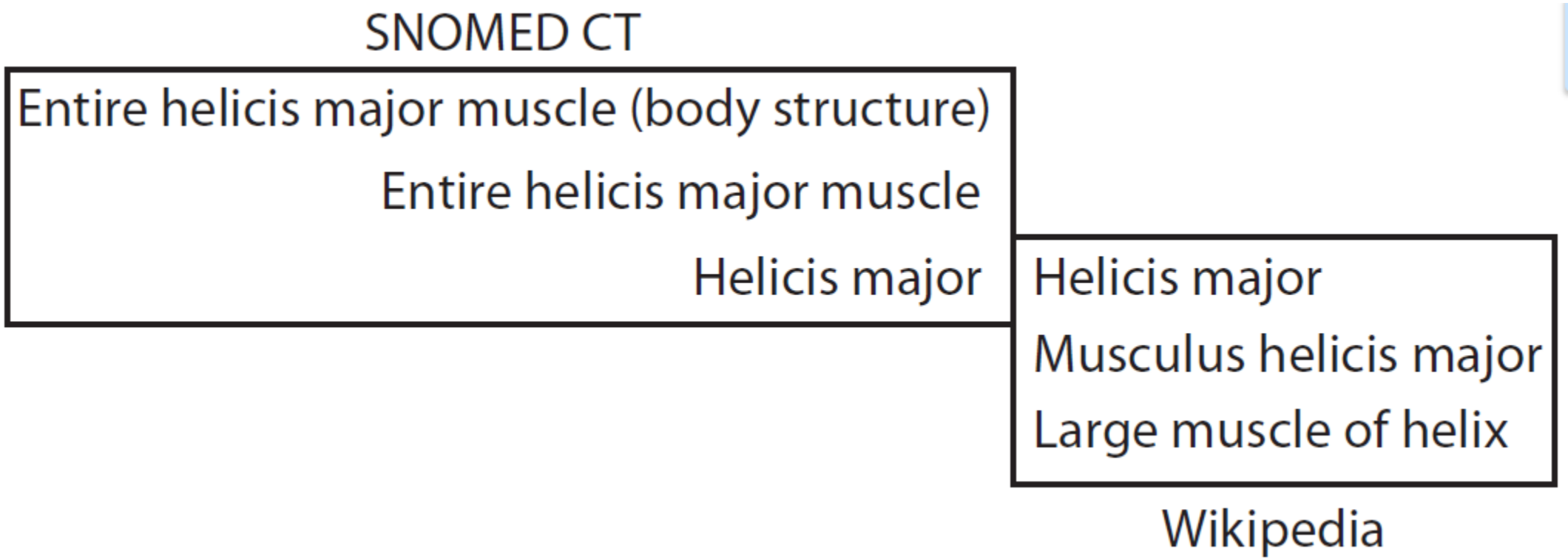
Wikipedia Redirects

- Wikipedia articles must have unique names
- There are “shadow” articles with no content which simply redirect to another article
  - Very often are synonyms of the article’s title

- Example: Heart attack redirects to myocardial infarction



Simple Matching Strategy



Initial Evaluation

- 43,580 exact matches between SNOMED CT and Wikipedia
  - 42,958 concepts had new synonyms from redirects
  - Extracted 446,053 new synonyms

- Random sample of 100 matches, consisting of 988 new synonyms
  - 407 synonyms (41.2%) were good
  - 360 synonyms (36.4%) were related, but incorrect
  - 221 synonyms (22.4%) completely unrelated

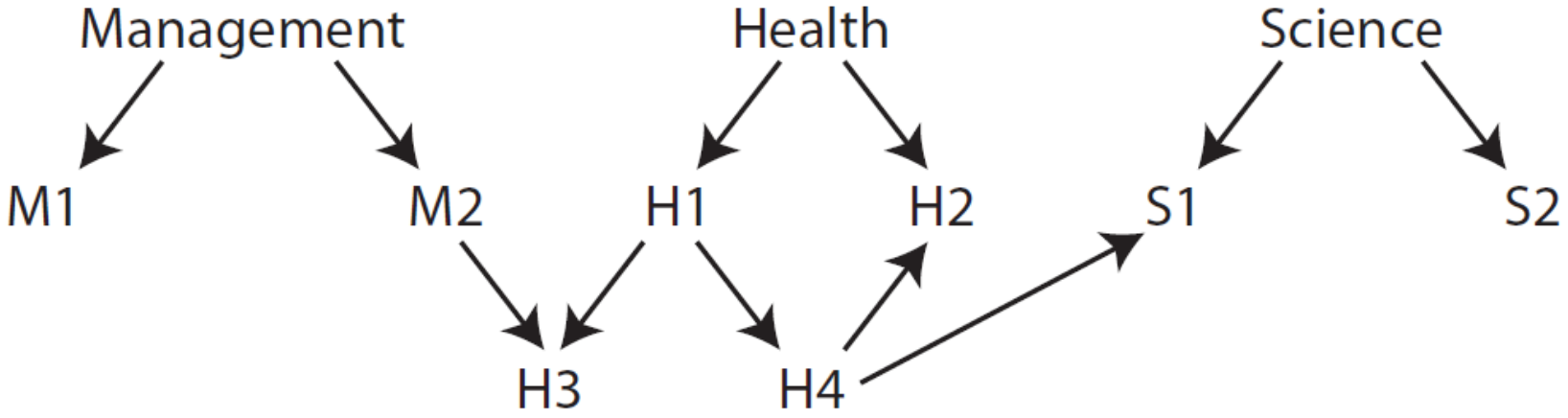
This isn't very good – we need to do better.

Strategy

- Understand the organization of Wikipedia better
  - Category hierarchy
- Analyze initial evaluation results
  - Classify matching errors
  - Look for solutions
- Implement solutions and re-evaluate

Wikipedia Categories

- Every Wikipedia page is a member of one or more categories
  - *E.g.,* “Health”
- Categories are organized into a graph
  - Cycles, cross-category links, multiple inheritance, etc…



  - We’ve done some heuristic pruning to make it usable

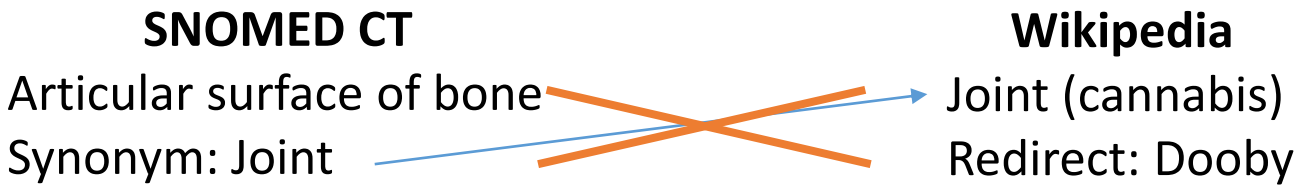
Two Strategies for Pruning

1. If the number of Wikipedia pages below a category is >900,000, then recursively look at subcategories up to depth 2, checking if the same is true for that category. Cut hierarchical links to stop this, either at depth 2 or earlier if possible.
2. Traverse Wikipedia from the top-level category breadth-first, giving each category a number for its depth. If a subcategory of a category has a depth lower than that category, cut the hierarchical link.

The results of these two strategies are combined.

Problem 1: Matches Outside Domain

Example:

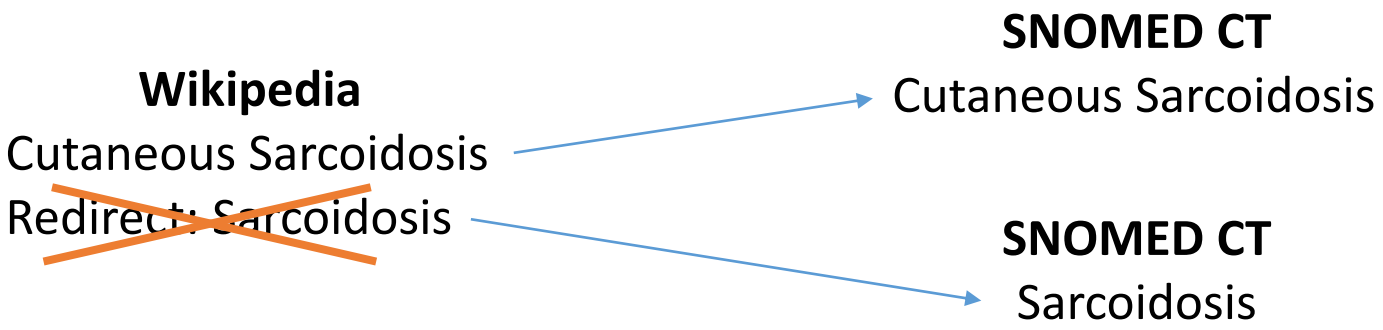


- Solution:
  - Use mapping of SNOMED CT semantic types to Wikipedia categories.
    - Require 50% of an article's categories to be in the SNOMED CT semantic type
    - Ensures domain is maintained (roughly)

SNOMED – Wikipedia Mapping

SNOMED CT Hierarchy	SNOMED CT Semantic Type	Wikipedia Categories
Body Structure	body structure	Anatomy
	cell structure	Cell anatomy
Clinical Finding	finding	Health
	disorder	Health
Geographical location / Environment	geographic location	Geography
	environment	Types of healthcare facilities, Buildings and structures, Human habitats
Event	event	Events
Observable entity	observable entity	Medical signs, Health care
Organism	organism	Organisms
Pharmaceutical / biologic product	product	Drugs, Proteins, Chemical substances, Body fluids
Physical force	physical force	Force, Physical quantities
Physical object	physical object	Physical objects
Procedure	procedure	Medical tests, Health care, Management
	regime/therapy	Medical treatments
Qualifier value	qualifier value	Articles
Record artifact	record artifact	Medical records, Documents, Technical communication
Social context	social concept	Human behavior, Society, Personal life
	ethic group	Ethnic groups
	racial group	Race (human classification)
Specimen	specimen	Biological specimens, Analytical chemistry
Staging and scales	tumor staging	Cancer staging
	staging scale	Medical scales, Cancer staging
	assessment scale	Medical scales
Substance	substance	Human proteins, Chemical substances

Problem 2: Incorrect, but related, redirects match other SNOMED terms



- Solution:
  - Eliminate redirects which match other SNOMED terms from the results.

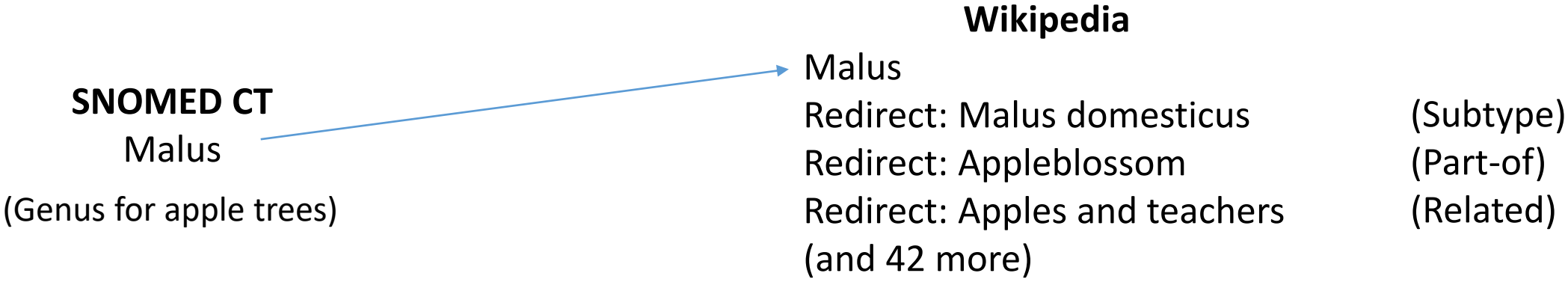
Problem 3: Acronyms are polysemous even within subdomains

Acronym “ED”

Eating Disorder
Effective Dose
Emergency Department
Erectile Dysfunction
…

- Solution
  - Don’t match acronyms to or from Wikipedia
  - Still allow acronyms to be learned as new synonyms

Problem 4: When there are large numbers of new synonyms, they are often unreliable



- Solution
  - Exclude synonym sets when they have >10 new synonyms.

Problem 5: Some subhierarchies in SNOMED aren't covered by Wikipedia (any matches are bad)

Example: Adjectives are not well covered by Wikipedia

- Solution:
  - Exclude some subhierarcies
    - Adjectival modifier
    - Specific site descriptor

Final Evaluation

- 30,781 SNOMED CT concepts matched with Wikipedia
  - 26,580 have new synonyms
  - 183,100 new synonyms added
- Evaluated 100 matches (517 synonyms)
  - 452 (85.6%) were correct matches
  - 76 (14.4%) were incorrect, but related
  - 1 (0.2%) were incorrect and unrelated

Match Details – Correct Results

- 61.95% were either word or term synonyms
- 18.14% were capitalization, spelling, or morphological variants
  - "Zinc sulfate (substance)” -> "Zinc Sulphate”
- 14.60% were various structured codings
  - "Calcium sulfate (product)” -> "CaSO4"
- 4.65% were shortened or extended forms
  - "Sedang language (qualifier value)” -> "Sedang"
- 0.44% were eponyms
  - "Chalcopsitta atra (organism)” -> "Bernstein's Black Lory"
- 0.22% were word order variants
  - "Gland of Zeis (body structure)” -> "Zeis' gland"

Match Details – Incorrect Results

- Incorrect but Related
  - 11.84% were subtypes
    - "Corydalis (organism)” -> "Corydalis adunca”
  - 24.36% were supertypes
    - “Congenital giant pigmented nevus of skin (disorder)” -> " Giant hairy nevus”
    - "New Zealand wren (organism)” -> "Xenicidae” (historic family name)

Match SNOMED Type Details

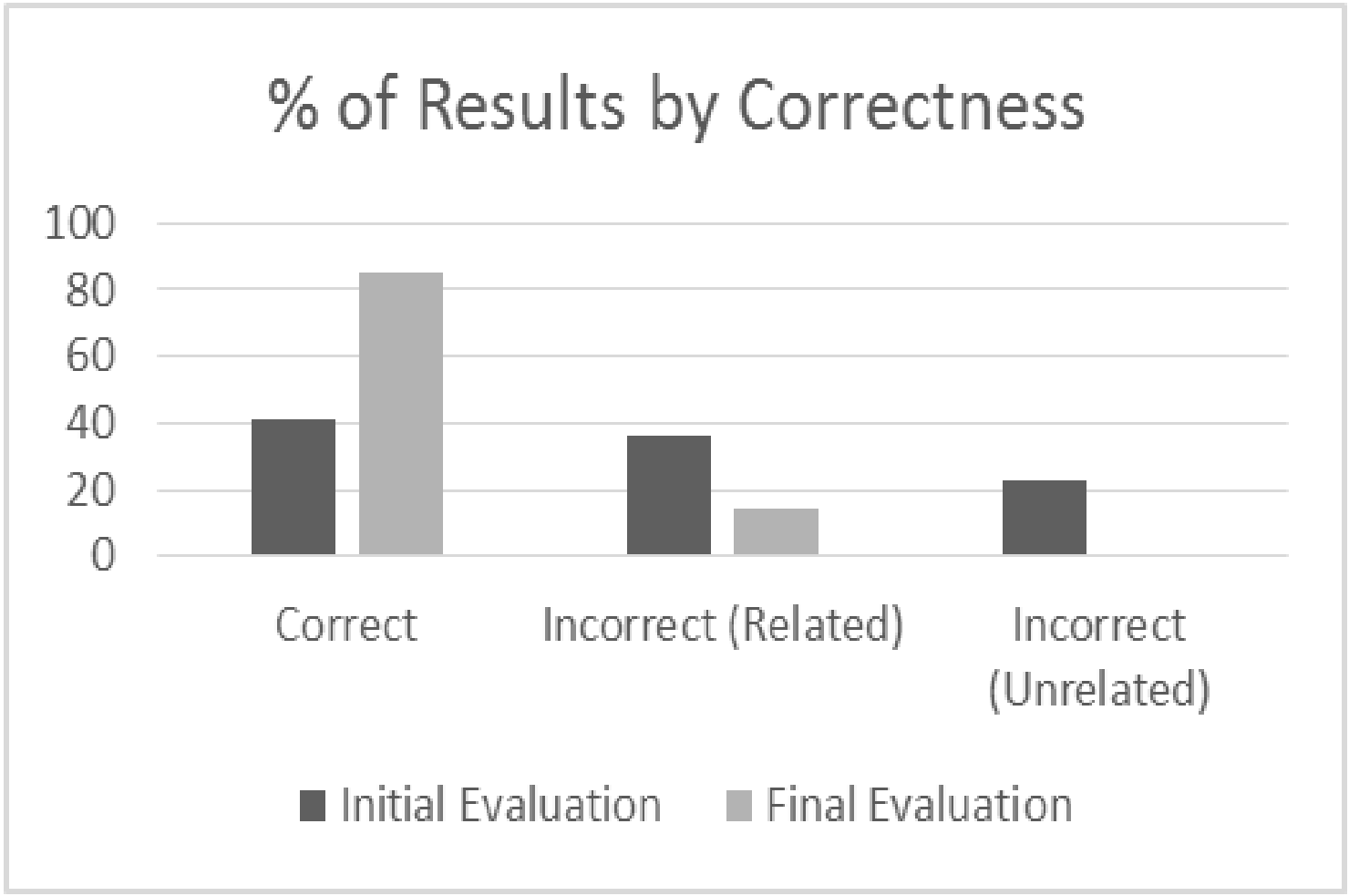
- Most matches are from the semantic types *body structure* (13.5%), *disorder* (17.9%), *organism* (22.4%), *product* (8.7%), and *substance* (26.4%).

Type	Match Count	Correct Syns.	Incorrect Related	Incorrect Unrelated
body structure	40	126	25	0
disorder	40	174	42	1
organism	40	117	1	1
product	38	628	39	0
substance	40	293	0	0
all others	40	151	85	1

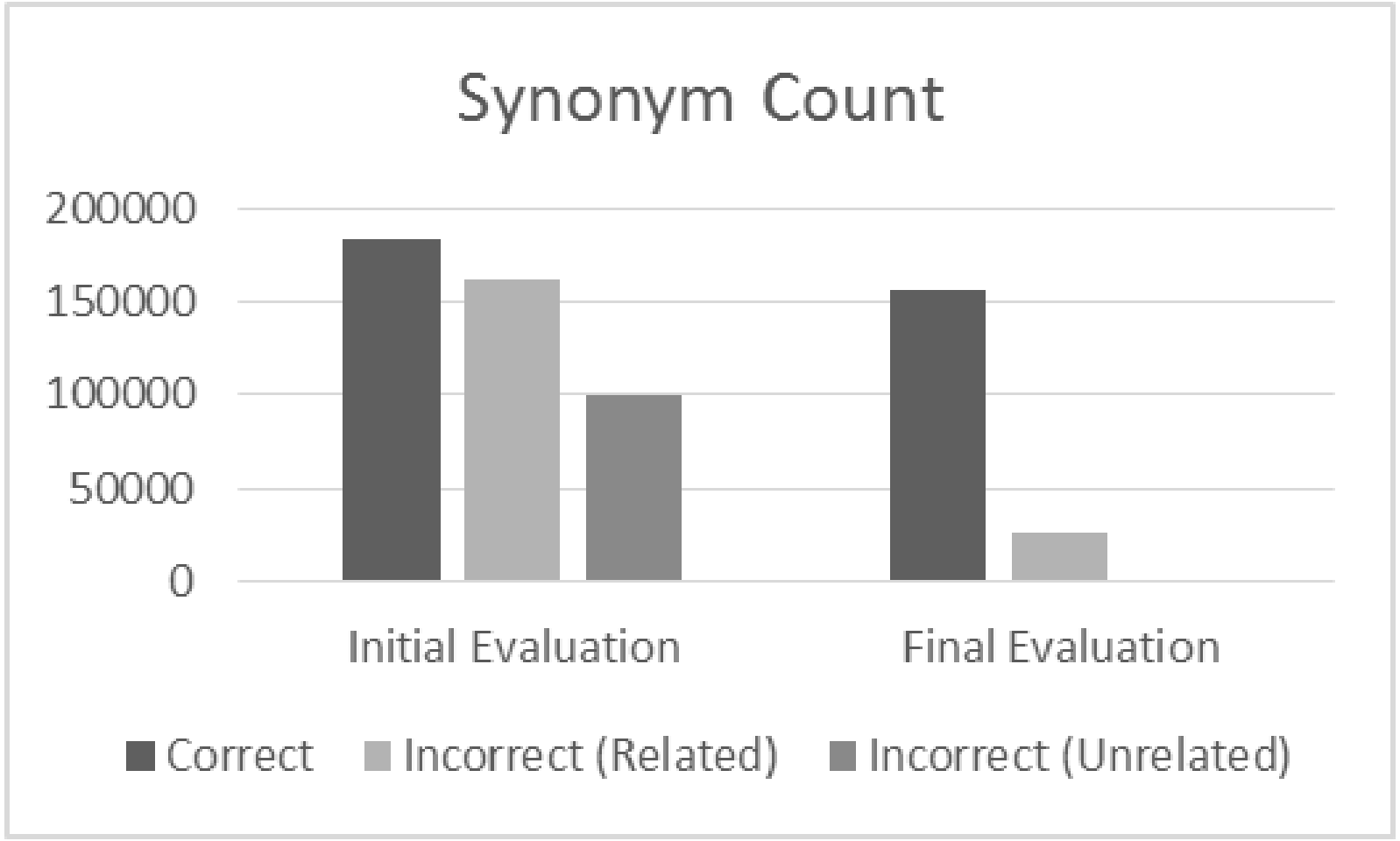
Reasons for Incorrect, but Related, Results

- Wikipedia contains redirects for non-synonymous terms, where the redirect name is just a small section of the overall article.
  - E.g., black vomit -> yellow fever
- Subtypes of Wikipedia page names redirect to a single page, where pages for those subtypes do not exist on their own.
  - E.g., *Diaptomus rostripes -> Diaptomus*
- SNOMED CT synonyms are sometimes more vague than the preferred term.
  - *Lower leg* has synonym *leg*

Comparing Initial and Final Evaluation



Comparing Initial and Final Evaluation



Availability

https://github.com/UBBiomedicalInformatics/WikiSNOMEDSynonyms

- JSON file
- We'd like community input
  - Bad synonyms
  - Missing synonyms
  - -> Will be rolled into future auto-generated versions

File a report on GitHub!

- As results improve, perhaps SNOMED will adopt our synonyms.

- Eventually we'd like to tag Wikipedia articles with SNOMED IDs.

Conclusion

- Examined issues in using Wikipedia as a source for synonyms
- Produced a high-level mapping between SNOMED CT semantic types and Wikipedia categories
- Increased number of synonyms in SNOMED CT by 183,100 with precision of 85.6%.

- Future work: Compare with UMLS sources

Thanks for Listening!

Questions?

https://github.com/UBBiomedicalInformatics/WikiSNOMEDSynonyms