MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-619

Automatically Expanding the Synonym Set of SNOMED CT using Wikipedia

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

Department of Biomedical Informatics, University at Buffalo, SUNY, Buffalo, NY, USA

Abstract

Clinical terminologies and ontologies are often used in natural language processing/understanding tasks as a method for semantically tagging text. One ontology commonly used for this task is SNOMED CT. Natural language is rich and varied: many different combinations of words may be used to express the same idea. It is therefore essential that ontologies and terminologies have a rich set of synonyms. One source of synonyms is Wikipedia. We examine methods for aligning concepts in SNOMED CT with articles in Wikipedia so that newly-found synonyms may be added to SNOMED CT. Our experiments show promising results and provide guidance to researchers who wish to use Wikipedia for similar tasks.

Keywords:

Ontology; Natural Language Processing; Terminology

Introduction

Automated understanding of text within the medical domain relies heavily upon the coverage of clinical terminologies. One such terminology, SNOMED CT, has been used extensively for such tasks [1-10]. An observation which has been noted by researchers examining SNOMED CT's coverage of clinical problem lists is that it could benefit from a more expansive set of synonyms [11].

As of the March 2014 English release, SNOMED CT contains 403,465 concepts, organized into several hierarchies covering medicine and medicine-related domains. Each of the 403,465 concepts contains at least two textual representations – one which includes one of a few dozen semantic types for disambiguation purposes, and one without. To these, there are added 230,863 synonyms which generally consist of alternative names, abbreviations, and shortened forms.

Wikipedia [12] is a community-maintained encyclopedia, covering topics in nearly every imaginable domain. It has a large number of articles related to medicine and science [13], and its scientific articles are of similar quality as Encyclopedia Brittanica [14]. Subdomains of its medicine-related articles, such as mental health, have been shown to have similar accuracy when compared to curated web sources [15]. As of December 2014,[1] Wikipedia contained over 4.6 million articles in English. One source of synonyms in Wikipedia is page redirects.

Wikipedia is being used increasingly often in medicine. In certain subdomains it has shown to be useful as a patient education resource [16]. In 2009, 28% of pharmacists reported using Wikipedia for drug information [17]. It has also been shown to be useful in monitoring infectious disease [18]. Wikipedia is being used in medical research with increasing frequency. Google Scholar finds about 20,300 results for the terms "wikipedia" and "medical" since 2013, and only an extra 1,400 since 2010.

Every article in Wikipedia is tagged with one or more of over 900,000 categories, which form a directed graph ("hierarchy"). Unfortunately, Wikipedia's "hierarchy" of categories "is barely useful for ontological purposes" [19]. For example, *Cath lab* is a page in the category *Cardiac procedures*. A cath lab is certainly a *place* where a cardiac procedure may be performed, but it is not itself a cardiac procedure. Indeed the categorization of pages is often more similar to a collection of related topics, rather than a rigorous ontological classification. For this reason, the ontology alignment techniques which have previously been used with SNOMED CT (such as [20, 21]) are not helpful.

We develop a method for automatically matching SNOMED CT concepts to Wikipedia articles based on lexical matches between synonyms from both, using heuristics, and through alignment of useful portions of the Wikipedia category hierarchy with SNOMED CT semantic types. Previous research has categorized Wikipedia articles as being either health-and-clinically related, or not, using the Wikipedia category hierarchy [22].

Elkin et al. [8] have used a set of 2.5 million synonyms created through a knowledge engineering process in the iNLP system. They used word synonyms such as *cancer* for *neoplasm*, then propagated the synonym through all concepts using the original word, creating new term synonyms. So, the synonym *cancer of tonsil* is added for *neoplasm of tonsil*. One issue with this approach is that many of the synonyms may never appear in actual text. Wikipedia redirects, on the other hand, are created because they are believed likely to occur. We are, of course, not the first to extract synonyms from Wikipedia (see [23] for one of the earliest examples), but we believe we are the first to attempt to enhance SNOMED CT synonymy using heuristic based extraction methods with Wikipedia.

Methods

Wikipedia's redirect pages (henceforth, *redirects*) have no content; they only automatically redirect the user to a specific page. Redirects are designed to get the user to the most appropriate page given their search. Redirects may be: alternative names; plurals; closely related words; pointers from adjectives/adverbs to the noun form; less or more specific forms of names; abbreviations; alternative spellings and common misspellings; alternately punctuated forms; alternative capitalizations; and subtopics within a larger article

[1] We use version 20140614 in this study.

[24]. Page redirects appear to be the most common way to derive synonymy from Wikipedia.

There are other methods for deriving synonymy using Wikipedia. Bolded words in an article's lead section are often synonyms according to the Wikipedia Manual of Style [25]. Analysis of links between pages, and more complex linguistic analysis of the page may also be used. In this paper we focus on redirects, but recognize that the other options may be useful, and leave analysis of them to future work.

SNOMED CT concepts are matched with Wikipedia articles through lexical matches. For example, in Figure 1 you can see that SNOMED CT contains the concept *Entire helicis major muscle (body structure)* with two synonyms. One of those synonyms, *Helicis major*, is also in Wikipedia. Wikipedia then contributes two new synonyms: *Musculus helicis major* and *Large muscle of helix*.

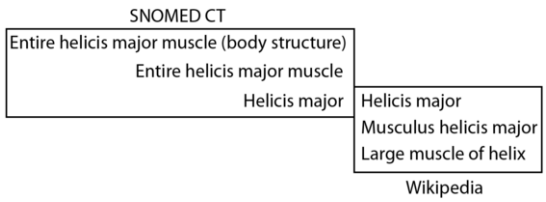

Figure 1 - A match between SNOMED CT and Wikipedia

We first performed an initial evaluation using the naïve matching strategy described above, generating the set of matches between SNOMED CT concepts and Wikipedia articles. We found that simply matching synonyms from SNOMED CT and Wikipedia does not produce very good results (see the Results section). However, analysis of the initial results led to the development of several heuristics, including a mapping between SNOMED CT semantic types and Wikipedia category "hierarchies". We hypothesized the heuristics and the mapping would improve the results.

In the remainder of this section, we discuss some preliminaries about Wikipedia's category hierarchy and what methods we used to make use of it. Then, the problems found with the initial evaluation are presented and solutions proposed. Finally methodology we used in the final evaluation is presented.

Wikipedia Categories

Every page in Wikipedia is a member of one or more categories, defined by the community according to the categorization guidelines [26]. Categories are organized into a directed graph which contains cycles, and includes edges between "hierarchies" from higher to lower level categories. The category graph also allows for multiple inheritance: a subcategory may have more than one supercategory. This structure is illustrated in Figure 2. These characteristics are so prevalent that often the closure of subcategories of an upper level category is all or most of Wikipedia.

To overcome these difficulties with the category graph, we use two independent approaches, with the results later combined. The first approach is naïve – it simply determines if the subcategorical closure of a category is all or most of the categories in Wikipedia. If it is, it recursively navigates up to *d* levels below the initial category, at each step checking again if the subcategorical closure is all or most of the categories in Wikipedia. At the point where the subcategorical closure is not all or most of Wikipedia, or depth *d* is reached, the algorithm cuts subcategory links. For this study, we used *d*=2.

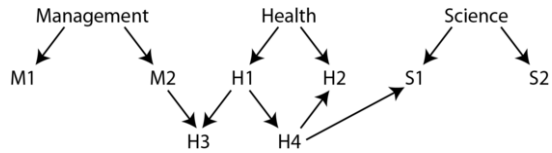

Figure 2 - An illustration of the structure of the Wikipedia category "hierarchy".

The second approach we use is similar to that of [22]: using a breadth-first approach we traverse the graph down from the top-level category (*Articles*) assigning each category a number for its depth. If a category has a subcategory which has a depth less than that of the category itself, that relation is removed.

Alignment

During careful analysis of the initial results, we determined that poor matches and synonyms occurred for five reasons. For each of these we have proposed a solution.

1. Problem: Many incorrect matches between SNOMED CT and Wikipedia (those where none of the resulting synonyms are relevant) occur because a SNOMED CT synonym matches with a Wikipedia redirect for a page outside the appropriate domain.
   Example: The SNOMED CT concept *articular surface of bone* has the synonym *joint*. In Wikipedia, *joint* is polysemous. One use is for drugs, where synonyms such as *dooby* are found.
   Solution: We use the mapping in Table 1 to ensure the domain is maintained. Only the included semantic types are matched on. We require 50% of an article's categories to be in the SNOMED CT semantic type of choice. This value was chosen based on work by [22] categorizing health data in Wikipedia.
2. Problem: Related, but incorrect, redirects often are exact matches of other SNOMED CT terms.
   Example: In Wikipedia, *cutaneous sarcoidosis* redirects to the article *sarcoidosis*. In SNOMED CT, these are two distinct (but related) concepts.
   Solution: Eliminate redirects which match other SNOMED terms from the results.
3. Problem: Acronyms are very polysemous, even within subdomains.
   Example: The acronym *ED* can stand for: eating disorder, effective dose, emergency department, erectile dysfunction, and others.
   Solution: Acronyms are excluded from match criteria, but not results.
4. Problem: If there are more than 10 new synonyms found for SNOMED CT terms outside substances, products, disorders, and observable entities, the new synonyms are often unreliable.
   Example: The SNOMED CT concept *Malus* (the genus for apple trees) when matched against Wikipedia results in 47 new "synonyms", many of which are subtypes like *Malus domesticus*, parts such as *Appleblossom* or even related topics like *apples and teachers*. Out of the 47, only 5 were actual synonyms.
   Solution: These are removed from the results.

Table 1 - Mapping between SNOMED CT semantic types and Wikipedia categories

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
	ethnic 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

5. Problem: Some subhierarchies of SNOMED CT semantic types contain data not within Wikipedia, and any matches will likely be incorrect.
   Example: The subhierarchy *adjectival modifier* below *qualifier value* contains many adjectives, while adjectives are not well covered by Wikipedia
   Solution: The subhierarchies *adjectival modifier*, and *specific site descriptor* are excluded.

Final Evaluation Methodology

Evaluation was performed again after the solutions in the Alignment section were applied. Two researchers, DRS and PLE, independently created the ground truth used in this study. One-hundred matches and all of their resulting synonyms were randomly sampled and scored. Both annotators classified each new synonym as either being: correct; incorrect, but a subtype; incorrect, but a supertype; incorrect, but related otherwise; or incorrect and unrelated. If a synonym was incorrect but also would never occur in the domain it was excluded from the evaluation results. If a synonym was correct, annotators would classify the synonym as one of: morphological variant – those which a stemmer would find equivalent; spelling variant; capitalization variant; shortened or extended form; eponym; structured coding; or word or term synonym. The number of correct and incorrect synonyms found by DRS and PLE were compared using Cohen's kappa coefficient to measure inter-annotator agreement (κ=.77). Discrepancies between the two annotation results were examined by DRS, and the most correct annotation result was accepted. If a most correct option was not obvious, DRS and PLE discussed the discrepancy until a consensus could be reached.

Results

Initial Evaluation

In this trial we found 43,580 exact lexical matches between SNOMED CT and Wikipedia with 42,958 of those having new synonyms. From these we extracted 446,053 new synonyms. We sampled 100 of the matches (consisting of 988 new synonyms). A single researcher examined these results carefully and found that only 407 (41.2%) of the new synonyms were valid. An additional 360 synonyms (36.4%) were related to the SNOMED CT concept, but were incorrect. These were often related to a higher or lower level concept. The remaining 221 (22.4%) were completely unrelated.

Final Evaluation

After elimination of SNOMED CT concepts from semantic types we do not believe are covered well by Wikipedia (those not in Table 1), there are 272,613 SNOMED CT concepts. Using the heuristics and matching techniques detailed in the methodology section, our system matches 30,781 SNOMED CT concepts. Of those which are matched, 26,580 have new synonyms. Out of the box, SNOMED CT contains 230,863 synonyms. To those, we add an additional 183,100.

Of the 517 synonyms analyzed, we found that 452 (85.6%) of them were correct (true positives), 76 (14.4%) of them were related but incorrect, and 1 (0.2%) was incorrect and unrelated. These percentages are significantly better than the initial evaluation, as visualized in Figure 3.

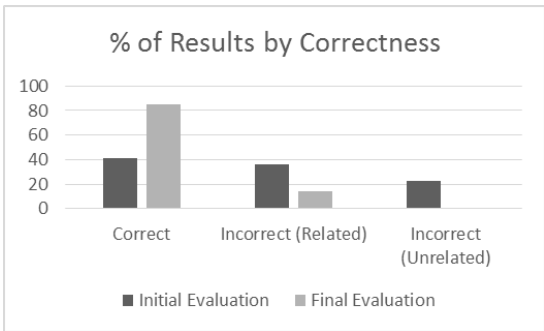

Figure 3 –Percent of correct, incorrect but related, and incorrect and unrelated results in the initial and final evaluations.

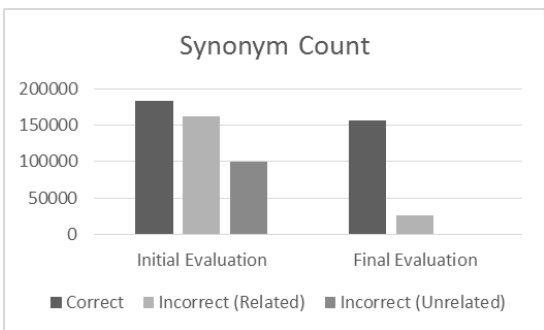

Figure 4 –Correct, incorrect but related, and incorrect but unrelated results in the initial and final evaluations.

Not only did precision improve as a result of our matching technique, few of the correct synonyms found in the initial evaluation were incorrectly eliminated (Figure 4). If we take the sample percentages and apply them to the total number of found synonyms, then compare the initial to final evaluations, we find that the number of correct synonyms decreased from 183,748 to 156,744 (14.7% change), incorrect but related decreased from 162,529 to 26,305 (83.82%), and incorrect and unrelated decreased from 99,775 to 346 (99.65%).

Of the correct results, we found that 61.95% were either word or term synonyms, 4.65% were shortened or extended forms, 0.44% were eponyms, 0.22% were word order variants, 14.60% were various structured codings, and the final 18.14% were capitalization, spelling, or morphological variants. Of the incorrect but related results, we found that 11.84% were subtypes and an additional 24.36% were supertypes. Very few of the synonyms we evaluated would never occur in the domain – only 7 of the total evaluated (1.3%).

Most matches are from the semantic types *body structure* (13.5%), *disorder* (17.9%), *organism* (22.4%), *product* (8.7%), and *substance* (26.4%). Our 100 sampled matches closely followed these percentages. In order to better understand the characteristics of the matching algorithm in these categories, we had an annotator examine random additional examples from these categories until we had around 40 matches from each (see Table 2).

The relative lack of unrelated incorrect results in this table shows that our matching method is correctly matching SNOMED CT terms with Wikipedia articles for the exact concept or a closely related concept.

Table 2 - Statistics for common semantic types

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

Comparison with Previous Work

We compared the synonyms created using our process with those from Elkin et al.'s iNLP system. We found that only 8,222 of our new synonyms were in their expanded synonym set. We did not use a knowledge engineering approach as they did, which likely would have inflated this number greatly.

Discussion

Using community-sourced data is difficult, because the level of rigor used in its creation is not always of the highest level. The methods we've discussed do not give perfect results, though they are much better than a naïve approach. To help understand what made our system imperfect, we have conducted error analysis on incorrect synonyms.

We have identified three major classes of reasons for incorrect but related synonyms. First, in some cases Wikipedia has redirects from the looked up SNOMED CT term to a related article in which discussion of the concept is only a small part of the article. For example, symptoms of a disease might redirect to a closely associated disease. Consider that *black vomit* redirects to *yellow fever*, since vomit containing blood is a major symptom of yellow fever (and in Spanish, yellow fever is known as *vomito negro* for this reason).

The second reason for incorrect but related synonyms is that a mid-level SNOMED CT concept matches a Wikipedia page which has redirects from more specific or closely associated terms or topics without pages themselves. For example, the page for the genus *Diaptomus* has a redirect from *Diaptomus rostripes* which is a species in that genus with no page itself.

The final reason is that SNOMED CT sometimes has synonyms which are more vague than the concept they stand for. For example, SNOMED CT has a concept for *lower leg*, which is usually defined as the lower extremity from the knee, to the ankle. One of the synonyms is simply *leg*. This is not really a synonym, and that it matches with a more general Wikipedia article should not be surprising.

Incorrect and unrelated terms occurred generally for two reasons. In some cases Wikipedia simply had incorrect redirects. For example *SkyUnion* redirected to *Immunoglobulin G*, while *Sky Union* appears to be a video game company. Some semantic types in SNOMED CT are extremely broad, most notably *qualifier value*. There are many subhierarchies of *qualifier value* which match Wikipedia very well, but some do not. We made an effort to eliminate bad subhierarchies which contained mostly adjectives, but there are others which are problematic. Specifically, it seems that matches from the concepts under *context values* are particularly low quality (*e.g.*, the concept *Done* matches a Wikipedia page about an album named *Done)*. In very broad categories, our matching heuristics have only a small impact.

Some Wikipedia categories such as *Organisms* and *Chemical substances* seem to have articles with very high quality redirects. Chemical substances in particular often have articles for very specific substances instead of redirecting to classes of substances, even if the specific article has little text. Many articles in these categories include codings such as enzyme codes and molecular formulas.

Future research should explore methods to expose deep semantic understanding of articles to both extract more synonyms and ensure correctness. We believe deeper understanding of Wikipedia articles, combined with using parents and children of a SNOMED CT concept being matched, may improve our results. We will also explore methods for attaching provenance and confidence to synonyms. Eventually it is our goal to release frequently updated versions of our synonym set publicly,[2] and use crowdsourcing techniques to continually increase quality.

Conclusion

It is well known that the more synonyms which are available for terminological concepts, the more easily ontologies and terminologies may be used for natural language processing and understanding tasks. We have used Wikipedia redirects as a source to increase the number of synonyms in SNOMED CT by 183,100 with precision of 85.6%. Our techniques for matching SNOMED CT concepts against Wikipedia articles have produced a significant improvement over naïve approaches. Moreover, our experiences with using Wikipedia in this research project may be a valuable resource for other researchers looking to use Wikipedia as an enrichment source.

Acknowledgements

The authors would like to acknowledge Matt Hudson for his helpful comments on previous drafts of this paper.

References

[1] Brown, S.H., et al. eQuality for all - extending automated quality measurement from free text clinical narratives. AMIA Annu Symp Proc. 2008:71-5.

[2] Elkin, P.L., et al., Comparison of NLP Biosurveillance Methods for Identifying Influenza from Encounter Notes. Ann Intern Med, 2012. 156(1 Pt 1): 11-8.

[3] Elkin, P.L., et al. Aequus communis sententia: defining levels of interoperability. Stud Health Technol Inform. 2007;129(Pt 1):725-9

[4] Elkin, P.L., et al. NLP-based identification of pneumonia cases from free-text radiological reports. AMIA Annu Symp Proc. 2008:172-6.

[5] Elkin, P.L., et al., Secondary use of clinical data. Stud Health Technol Inform, 2010. 155: p. 14-29.

[6] Fitzhenry, F., et al., Exploring the fronteir of electronic health record surveillance: the case of postoperative complications. Med Care, 2013. 51(6): p. 509-516.

[7] Garvin, J.H., et al., Automated Quality Measurement in Department of the Veterans Affairs Discharge Instructions for Patients with Congestive Heart Failure. J Healtc Qual, 2012.

[8] Matheny, M.E., et al., Detection of infectious symptoms from VA emergency department and primary care clinical documentation. Int J of Med Inform, 2012. 81(3): p. 143-156.

[9] Matheny, M.E., et al. Detection of blood culture bacterial contamination using natural language processing. in AMIA Annu Symp Proc. 2009: 411-5.

[10] Murff, H.J., et al., Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA, 2011. 306(8): p. 848-855.

[11] Elkin, P.L., et al., Evaluation of the Content Coverage of SNOMED CT: Ability of SNOMED Clinical Terms to Represent Clinical Problem Lists. Mayo Clinic Proc, 2006. 81(6): p. 741-748.

[12] Wikipedia:About - Wikipedia, the free encyclopedia. [cited 2014 Dec. 6]; Available from: http://en.wikipedia.org/wiki/Wikipedia:About.

[13] Friedlin, J. and C.J. McDonald, An evaluation of medical knowledge contained in Wikipedia and its use in the LOINC database. J Am Med Inform Assoc, 2010. 17(283e287): p. 283e287.

[14] Giles, J., Internet encyclopaedias go head to head. Nature, 2005. 438(7070): p. 900-901.

[15] Reavley, N.J., et al., Quality of information sources about mental disorders: a comparison of Wikipedia with centrally controlled web and printed sources. Psychological medicine, 2012. 42(08): p. 1753-1762.

[16] Thomas, G.R., et al. An evaluation of Wikipedia as a resource for patient education in nephrology. in Seminars in dialysis. 2013. Wiley Online Library.

[17] Brokowski, L., et al, Evaluation of pharmacist use and perception of Wikipedia as a drug information resource. Ann Pharmacother, 2009. 43(11): p. 1912-1913.

[18] Generous, N., et al., Global disease monitoring and forecasting with wikipedia. PLoS Comput Biol, 2014. 10(11): p. e1003892.

[19] Suchanek, F.M., G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. in Proceedings of the 16th international conference on World Wide Web. 2007..

[20] Zhang, S., et al. Alignment of multiple ontologies of anatomy: deriving indirect mappings from direct mappings to a reference. AMIA Annu Symp Proc. 2005.

[21] Bodenreider, O., Comparing the representation of anatomy in the FMA and SNOMED CT. AMIA Annu Symp Proc. 2006.

[22] Liu, F., et al. Automatically identifying health-and clinical-related content in wikipedia. in MedInfo. 2013.

[23] Milne, D., O. Medelyan, and I.H. Witten. Mining domain-specific thesauri from wikipedia: A case study. in Proceedings of the 2006 IEEE/WIC/ACM international conference on web intelligence. 2006..

[24] Wikipedia:Redirect - Wikipedia, the free encyclopedia. [cited 2014 Dec. 9]; Available from: http://en.wikipedia.org/wiki/Wikipedia:Redirect.

[25] Wikipedia:Manual of Style/Text formatting. [cited 2014 Dec. 9]; Available from: http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Text_formatting.

[26] Wikipedia:Categorization - Wikipedia, the free encyclopedia. [cited 2014 Dec. 9]; Available from: http://en.wikipedia.org/wiki/Wikipedia:Categorization.

Address for correspondence

Daniel R. Schlegel, PhD, Department of Biomedical Informatics, University at Buffalo, 923 Main Street, Buffalo, NY 14203

[2] The synonym set is available for download in the Research section of the UB Biomedical Informatics website: http://www.smbs.buffalo.edu/biomedicalinformatics/.