Title

Cross-domain targeted ontology subsets for annotation: The case of SNOMED CORE and RxNorm

Authors

López-García, Pablo
LePendu, Paea
Musen, Mark
et al.

Publication Date

2014-02-01

DOI

10.1016/j.jbi.2013.09.011

Peer reviewed

Published in final edited form as:
J Biomed Inform. 2014 February ; 47: 105–111. doi:10.1016/j.jbi.2013.09.011.

Cross-domain targeted ontology subsets for annotation: The case of SNOMED CORE and RxNorm

Pablo López-García[a,b,*], **Paea LePendu**[a], **Mark Musen**[a], and **Arantza Illarramendi**[b]

[a]Stanford Center for Biomedical Informatics Research, Stanford University, Medical School Office Building, Room X-215, 1265 Welch Road, Stanford, CA 94305-5479, USA

[b]Department of Computer Languages and Systems, University of the Basque Country UPV/EHU, Manuel de Lardizabal 1, 20018 Donostia-San Sebastián, Spain

Abstract

The benefits of using ontology subsets versus full ontologies are well-documented for many applications. In this study, we propose an efficient subset extraction approach for a domain using a biomedical ontology repository with mappings, a cross-ontology, and a source subset from a related domain. As a case study, we extracted a subset of drugs from RxNorm using the UMLS Metathesaurus, the NDF-RT cross-ontology, and the CORE problem list subset of SNOMED CT. The extracted subset, which we termed RxNorm/CORE, was 4% the size of the full RxNorm (0.4% when considering ingredients only). For evaluation, we used CORE and RxNorm/CORE as thesauri for the annotation of clinical documents and compared their performance to that of their respective full ontologies (i.e., SNOMED CT and RxNorm). The wide range in recall of both CORE (29–69%) and RxNorm/CORE (21–35%) suggests that more quantitative research is needed to assess the benefits of using ontology subsets as thesauri in annotation applications. Our approach to subset extraction, however, opens a door to help create other types of clinically useful domain specific subsets and acts as an alternative in scenarios where well-established subset extraction techniques might suffer from difficulties or cannot be applied.

Keywords

Ontologies; SNOMED CT; RxNorm; NDF-RT; UMLS; Medical records; Annotation

1. Introduction

Biomedical ontologies are key to medical informatics, but their size and complexity still represent a challenge in many applications [1]. The Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) [2], for example, comprises over a million terms (text strings) structured as a taxonomy of 400,000 concepts and an ontological layer that conforms to the $\mathcal{EL}$ + + description logic standard [3]. Finding a portion of interest that can be used as a virtual substitute for a whole ontology for a specific application or domain is a highly desired objective, because it reduces complexity, improves maintenance, encourages

*Corresponding author at: Stanford Center for Biomedical Informatics Research, Stanford University, Medical School Office Building, Room X-215, 1265 Welch Road, Stanford, CA 94305-5479, USA. plopezgarcia@gmail.com, plopez@know-center.at (P. López-García).

reuse and customization, and improves performance in applications [4]. These portions of interest are referred to as subsets, modules, or segments.

An application where ontology subsets can play a major role is in the annotation of clinical documents. A key part of the annotation process consists of syntactically identifying ontology concepts in the free text of the document by using efficient string matching techniques and a reference thesaurus [5]. This strategy is followed, for example, by the National Center for Biomedical Ontology (NCBO) Annotator [6]. The content and size of the thesaurus play a key role in the annotation process, influencing which and how fast ontology concepts will be identified in free text.

The characteristics of the thesaurus are even more relevant in the case of annotators that use ontology matching and word-sense disambiguation techniques [7,8], such as MetaMap [9]. MetaMap can produce more accurate results than efficient string matching, but it is computationally much more expensive: Aronson and Lang showed that MetaMap can take up to a minute to process an average MEDLINE citation. Furthermore, they showed that some complex phrases can require hours of computation when using the 5 million terms from the Unified Medical Language System (UMLS) Metathesaurus [10] as a reference thesaurus, because hundreds of thousands of potential mappings are generated [9].

Several approaches for extracting compact subsets from ontologies that might be used as a thesaurus for annotation have been proposed. Research on ontology modularization, interested in modules that preserve the logical entailments that can be derived from the original ontology, has found efficient strategies of extraction by traversing an ontology from a set of input key concepts, or signature [4]. Ontology modularization techniques include graph-traversal [11–14] and logic-based techniques [15,16], whose extraction strategies depend on the ontology's topology and its definitional axioms, respectively. The requirement of preserving the original ontology entailments, key to ontology modularization, however, adds a large number of terms to the module that are unlikely to be found in clinical documents, necessarily affecting precision and performance [17].

Therefore, term-frequency analysis of large corpora or datasets is often preferred to produce small and precise subsets for annotation [18–20]. One of the most relevant examples of such a subset is the CORE problem list subset of SNOMED CT (CORE) [20], which is only 1.50% the size of SNOMED CT but covers over 90% of the diagnoses and problem lists found in existing reference datasets. Public authoritative medical corpora are very scarce (with the notable exception of the Multiparameter Intelligent Monitoring in Intensive Care II clinical database [21]), and using a generalist corpus (e.g., MEDLINE) might not provide good enough results because of potential mismatch between content and vocabulary used in scientific abstracts and clinical jargon [17]. To extract the CORE subset, seven large-scale health care institutions collaborated to analyze their datasets. It is expected that extracting a subset with similar characteristics as CORE for a different domain would require a comparable effort, which is unfeasible in many cases.

Furthermore, in some scenarios only terms from one or more specific domains might need to be identified in a medical document. Pharmacovigilance using clinical notes [22] and extraction of drug-disease treatment pairs from biomedical literature [23] are two representative examples. In these cases, using a thesaurus the size of the UMLS Metathesaurus or SNOMED CT imposes unnecesary overhead when annotating because only drugs and diseases are relevant in the free text. Domain-specific ontologies might be preferred or even required when annotating. In the United States, for example, SNOMED CT is the designated standard terminology for diagnoses and problem lists, but RxNorm is the standardized nomenclature for clinical drugs for use in federal government systems [24].

The already available CORE problem list subset of SNOMED CT represents a useful resource to annotate diseases, providing high recall while being exceptionally compact in comparison to SNOMED CT [20]. To the knowledge of the authors, no comparable subset of RxNorm is available to annotate drugs.

Burton et al. [25], however, showed that the National Drug File-Reference Terminology (NDF-RT) [26], a drug information source produced by the U.S. Department of Veteran Affairs, was extremely valuable for making inferences between medications and indications, because NDF-RT has comprehensive information on recommended treatments [27].

NDF-RT, SNOMED CT, and RxNorm are all included in the UMLS Metathesaurus [10], a biomedical ontology repository with mappings developed by the U.S. National Library of Medicine. Ontology mappings are links between concepts from different ontologies that are considered semantically equivalent and are the main field of study of ontology alignment or ontology matching [7]. The UMLS Metathesaurus not only contains NDF-RT, SNOMED CT, and RxNorm, but also mappings between them.

In this study, we explore the possibility of using mappings and cross-ontologies available in ontology repositories as an efficient way to extract compact subsets for annotation, given the fact that extremely compact high quality subsets such as CORE are already available. In particular, the aims of the present work are as follows.

1.1. Objectives

1. To propose an approach that extracts a target subset *TS* from a target domain ontology *TO*, using an existing, related source subset *SS* from a related source domain ontology *SO*. The approach uses a standard ontology repository, a cross-ontology linking *SO* and *TO*, and mappings, and requires no preexisting corpus or signature selection.
2. To study and compare the relative size and performance of the subsets *SS* and *TS* when used to annotate terms in clinical documents, as opposed to using their full domain ontologies *SO* and *TO*.

As a use case for evaluation, we extract a subset from RxNorm (*TO*) in the domain of drugs for treatment, which we term RxNorm/CORE (*TS*), using the existing CORE subset (*SS*) of SNOMED CT (*SO*) in the domain of diseases. We use the NDF-RT ontology as a link between *SO* and *TO*, and the UMLS Metathesaurus as an ontology repository that provides the ontologies and mappings between them. Finally, we study and compare the performance of RxNorm/CORE and CORE for identifying terms mentioned in medical research literature and discharge summaries.

2. Materials and methods

Fig. 1 shows an overview of our approach to obtain the RNorm/CORE subset, using the CORE problem list subset of SNOMED CT as source. The following subsections describe the approach in depth.

2.1. Materials

The Unified Medical Language System (UMLS) Metathesaurus [10] is a knowledge base that comprises over 160 biomedical ontologies (*source vocabularies* in UMLS terminology), including SNOMED CT, NDF-RT, and RxNorm. The UMLS Metathesaurus is part of the Unified Medical Language System, developed by the U.S. National Library of Medicine to facilitate interoperability between computer systems [28].

Terms that represent the same concept (e.g., 'heart attack', 'myocardial infarction', 'cardiac infarction', or 'infarction of heart') are assigned the same Concept Unique Identifier (CUI) in the UMLS Metathesaurus, regardless of which biomedical ontology they belong to. CUIs provide consistency for concepts and terms across ontologies, facilitating interoperability. The UMLS 2010AB release[1] was installed in a local MySQL database using MetamorphoSys, the UMLS installation and customization program. The UMLS Metathesaurus, comprising 158 source vocabularies in its 2010AB release, was accessed through standard SQL queries.

As an authoritative source subset for diseases, we selected the CORE problem list subset of SNOMED CT. The CORE subset is a subset containing 5814 concepts for documentation and encoding of clinical information at a summary level. The concepts included in the CORE subset represent the most frequently used terms in a series of datasets submitted by seven large-scale health care institutions that cover most medical specialties. The CORE subset provides a recall above 90% for diagnoses and problem lists with only 1.50% of the size of the full SNOMED CT [20]. Table 1 shows the five concepts in the CORE subset that were most frequently found in the submitted datasets.

Although the CORE subset is not part of the UMLS Metathesaurus, it is available online under the UMLS license. To maintain consistency, we used the 201102 version derived from UMLS Metathesaurus version 2010AB, which was the version of UMLS used throughout the study.

We selected the NDF-RT ontology to serve as the linking component. NDF-RT contains approximately 147,000 terms that represent 44,000 concepts, and it links our target and source domains (i.e., drugs and diseases). We were only interested in drugs used for treatment and we therefore used the relationship labeled as 'may treat'. The 'may treat' relationship indicates that "medication X is appropriate for the treatment of disease Y, its associated symptoms, or closely associated diseases" [26]. The remaining three relationships in NDF-RT that link both domains ('may prevent', 'may diagnose' and 'induces') were not used in this study.

Our target ontology was RxNorm, which is the standardized nomenclature for clinical drugs for use in U.S. federal government systems and which contains 437,000 terms that represent 194,000 concepts. The semantic approach used throughout the study follows the UMLS schema whereby two terms from the same or different ontologies were considered semantically equivalent if they shared the same CUI in the UMLS Metathesaurus.

2.2. Methods

The five steps that we followed to obtain the drugs in RxNorm related to diseases listed in the CORE subset using the UMLS Metathesaurus were as follows (see Fig. 1):

1. UMLS CUIs of diseases from the CORE subset were first identified.
2. Drug-disease pairs using the 'may treat' relationships in NDF-RT were extracted.
3. Identified NDF-RT diseases from step 2 were matched against CORE diseases from step 1.
4. Matching diseases identified at step 3 were used as a signature to follow 'may treat' relationships in NDF-RT and find related drugs.
5. Identified drugs in NDF-RT were finally matched against RxNorm.

[1]UMLS Metathesaurus 2010AB Release, http://www.nlm.nih.gov/pubs/techbull/nd10/nd10_umls.html.

The target subset, which we term RxNorm/CORE, consisted of drugs in RxNorm used to treat diseases in the CORE subset, as stated in the NDF-RT linking ontology.

2.3. Evaluation

Xu et al. [19] described a filtering approach to identify relevant concepts in UMLS by studying how many times each UMLS concept appeared in an external corpus, with encouraging results. A subsequent study confirmed the good results for SNOMED CT and showed that a filtering threshold higher than one (i.e., a concept is relevant if it appears at least twice in a document) severely affects precision and recall but only provides a marginal size reduction [17]. Therefore, in this study we considered a concept relevant for a corpus *C* if it appeared in at least one document of *C*. To measure the performance of a subset *S* (extracted from a domain ontology *O*) to annotate a corpus *C*, we define the relative size, precision, and recall of *S* as follows:

- $RelativeSize(S,O)=\frac{|Concepts(S)|}{|Concepts(O)|}$
- $RelativePrecision(S,O,C)=\frac{|RelevantConcepts(O,C)\cap Concepts(S)|}{|Concepts(S)|}$
- $RelativeRecall(S,O,C)=\frac{|RelevantConcepts(O,C)\cap Concepts(S)|}{|RelevantConcepts(O,C)|}$

For evaluation, we chose the following three heterogeneous corpora as annotation scenarios:

1. A subset of 200,000 records from MEDLINE, containing human case reports written in English from 2005 to 2010 [17].
2. The 27,000 discharge summaries available in the Multiparameter Intelligent Monitoring in Intensive Care II Research Database (MIMIC-II), which contained reasons for admission and previous conditions in the domain of diseases, as well as allergies, past medications, and indicated medications in the domain of drugs. The MIMIC-II database is a large collection of de-identified data from the Intensive Care Unit of Beth Israel Deaconess Medical Center offered to the research community [21].
3. A data set of 600 de-identified discharge summaries from Partners Healthcare, offered for the 2008 Natural Language Processing obesity challenge (NLP-O), that contains similar information to the one present in MIMIC-II [29].

Following the methodology described in López-García et al. [17] we built ranked versions of the domain ontologies (SNOMED CT and RxNorm), after storing the documents in the Lucene indexing engine. With this approach, all strings are normalized and exact string matching is used to match terms.

Three scores were added to each concept (MEDLINE, MIMIC-II, and NLP-O), representing the number of documents from each corpus where the concept had been identified (either in the title or abstract in the case of MEDLINE, and anywhere in the document in the case of the discharge summaries in MIMIC-II and NLP-O).

Our goal was to analyze the efficacy of using subsets for annotation and, more specifically, to determine if our method extracted a target subset (RxNorm/CORE) with similar efficacy to that of the source subset (i.e., CORE). Therefore, size, precision, and recall were calculated for both RxNorm/CORE and CORE itself, relative to their domain ontologies (i.e., RxNorm for RxNorm/CORE, and SNOMED CT for CORE). Because SNOMED CT is a comprehensive, multi-domain ontology, we did not use the full SNOMED CT as the reference domain ontology in our calculations when evaluating the CORE subset. Instead, we used the *clinical findings* and *diseases* hierarchies. Moreover, because we were only

interested in relative efficacy (i.e., what is missing when using a subset for annotation instead of a full ontology?), we did not make any assumptions about the absolute precision and recall of the full domain ontologies for annotating.

3. Results

Using the CORE subset's 5814 SNOMED CT concepts as input, we were able to extract a subset of 7499 related current drugs in RxNorm via NDF-RT's 'may treat' relationships and mappings in the UMLS Metathesaurus. Detailed results on the process, an analysis of CORE and the resulting RxNorm/CORE subset, and their performance in terms of relative size, precision, and recall are provided in the following subsections.

3.1. Obtaining the RxNorm/CORE subset

Fig. 2 summarizes the number of CUIs involved in each of the steps described in the Materials and Methods section.

1. The 5814 concepts in the CORE subset mapped to 5735 CUIs in UMLS.
2. Exactly 43,734 drug-disease pairs were identified in NDF-RT, which mapped to 34,930 CUI pairs in UMLS. 937 diseases, and 8400 drugs were involved.
3. From the previous 937 diseases, 553 were present in the CORE subset.
4. The previous 553 diseases were linked to 7755 drugs in NDF-RT via 'may treat' relationships.
5. The resulting 7755 NDF-RT drugs mapped to 7499 drugs in RxNorm.

3.2. The CORE subset, SNOMED CT, and UMLS

As shown in Fig. 2, the CORE subset's 5814 concepts mapped to 5735 UMLS CUIs (see Fig. 2). There was a direct one-to-one mapping between SNOMED CT IDs and UMLS CUIs for the large majority of concepts (97%). However, 51 UMLS CUIs mapped to 102 SNOMED CT concepts representing both a current and a concept to be replaced, the latter marked as to be retired from the subset.

In 28 cases, two concepts with different SNOMED CT IDs (e.g., 'Alcohol dependence' (ID 66590003) and 'Persistent alcohol abuse' (ID 284591009)) mapped to the same UMLS CUI (C0001973). In SNOMED CT, concepts in the diseases sub-hierarchy (ID 64572001, 63,884 concepts) represent 66% of the concepts in the clinical findings (ID 404684003, 96,783 concepts) hierarchy. In CORE, this figure increases to 83%, with 4133 diseases out of a total of 4968 clinical findings.

3.3. The RxNorm/CORE subset

The extracted RxNorm/CORE subset of 7449 concepts was 3.87% the size of the full RxNorm in terms of UMLS CUIs. The distribution of concepts according to RxNorm term types is shown in Table 2.

As displayed in Table 2, the majority of concepts in RxNorm/CORE corresponded to Semantic Clinical Drugs, which represent an ingredient plus strength and dose form. It must be noted, however, that all ingredients in Semantic Clinical Drugs are also present in RxNorm/CORE independently, because the 'may treat' relationship from NDF-RT links the diseases to indications in the form of ingredients as well. As an example, Table 3 shows an extract of the recommended treatments for edema (UMLS CUI C0013604), as captured by the 'may treat' relationship in NDF-RT. When considering only ingredients in RxNorm/CORE, the relative size of the subset is reduced to 0.4%.

3.4. Evaluation

The top part of Table 4 shows the relative size of RxNorm/CORE, and its performance in terms of precision and recall when used to annotate the MEDLINE, MIMIC-II, and NLP-O corpora. The size and performance were measured with respect to the domain ontology from which RxNorm/CORE was extracted (RxNorm). In the case of CORE, we used the *diseases* and *clinical findings* hierarchies of SNOMED CT (see Section 2.3).

The recall of RxNorm/CORE was below 35% for all corpora, although it was more compact with respect to its domain ontology. RxNorm/CORE showed better performance when used to annotate records from MEDLINE (34.79%) than discharge summaries from MIMIC-II (20.53%) and NLP-O (31.32%), the opposite of CORE. Fig. 3 provides a direct comparison between the SNOMED CT CORE subset and RxNorm/CORE with respect to their reference ontologies and domains for each corpus. The precision of CORE was significantly higher in all cases.

The exact number of documents for each corpus, the absolute size of the domain ontology and subsets, and the number of relevant concepts found are shown in Tables 5–7.

4. Discussion

Prior work has documented the effectiveness of using cross-ontologies and mappings to infer relevant concepts from related domains, as is the case with NDF-RT for drugs and diseases [25]. Our case study described a similar approach to obtain a compact subset of drugs from RxNorm (roughly 4% its size, 0.4% when considering ingredients only) using the CORE subset of SNOMED CT as source and the UMLS Metathesaurus to provide the mappings, which are both available to the biomedical informatics research community. Other available drug-indications linking components are the Medi-Span (Wolters Kluwer Health, Indianapolis) Drug-Indications Database,[2] and the following ones identified by Névéol and Lu [30]: MeSH,[3] DailyMed,[4] DrugBank,[5] and AHFS Consumer Medication Information.[6] These linking components, however, are not available in the UMLS Metathesaurus.

We followed Burton et al.'s approach based on their encouraging results and the accuracy of mappings between NDF-RT and RxNorm in the UMLS Metathesaurus [25]. However, our case study revealed that by using only direct mappings to infer related concepts from different ontologies, important information might be missed in some cases. Using UMLS CUIs, we were only able to directly map 553 (roughly 10%) of the diseases in the CORE subset to diseases in NDF-RT for which recommended treatment information existed. The most plausible explanation for this mismatch, in good agreement with Burton et al. [25], is that diseases in NDF-RT map to SNOMED CT terms that are more general than the ones found in problem lists, such as the CORE subset we used as a signature for the inferences. Problems when matching diseases between CORE and NDF-RT suggest that simple CUI matching between ontologies might not be an acceptable alternative in many cases, such as when the granularity of ontologies or of term usage is different. We also identified 28 cases in the CORE subset where two concepts with different SNOMED CT IDs mapped to the same UMLS CUI, which can be regarded as a semantic mismatch that should be taken into account when working with CUIs.

[2] http://www.medispan.com/drug-indications-database.aspx
[3] http://www.nlm.nih.gov/pubs/factsheets/mesh.html
[4] http://dailymed.nlm.nih.gov
[5] http://www.drugbank.ca
[6] http://www.ahfsdruginformation.com/products_services/ahfs_cmi.aspx

The limited precision and recall of RxNorm/CORE (Table 4) to annotate clinical documents confirms that extracting subsets for annotation is particularly challenging and that frequency-based techniques that use a clinical corpus close to the application are preferred if available [18,17]. In our case study, the performance of RxNorm/CORE to annotate drugs with respect to RxNorm was slightly better than using CORE to annotate diseases with respect to the diseases and clinical finding hierarchies of SNOMED CT in the case of MEDLINE. It must be noted that CORE is derived from clinical data, so it was expected to perform better in clinical corpora. Another issue that should be considered is the nature of the RxNorm/CORE subset that we extracted. Semantic Clinical Drug comprises the majority of the subset, but the size of the subset can be reduced an order of magnitude if only ingredients are to be identified in clinical documents, dramatically improving precision without sacrificing recall.

The qualitative benefits of using ontology subsets versus full ontologies in applications are many, and they are well-documented (complexity reduction, maintenance improvement, easy reuse, performance gain, etc.) [4]. However, the limited recall of both CORE (29–69%) and RxNorm/CORE (21–35%) in our experiments suggest that the benefits of using ontology subsets as thesauri in annotation applications should be reassessed.

5. Conclusions and future work

In this study, we have shown that cross-ontologies and biomedical ontology repositories with mappings are valuable tools to extract an ontology subset efficiently when another subset from a related domain is already available. In our case study, we used ontology subsets to serve as a reference thesaurus for annotating diseases and drugs in clinical documents. The CORE subset of SNOMED CT not only proved to be a useful resource to annotate diseases, but it also served as an authoritative source subset for extracting a related subset of drugs of RxNorm using our approach.

Our approach to subset extraction opens a door to help create other types of clinically useful domain-specific subsets, for example, in the domain of anatomy by using CORE as the source subset and the ‘finding site’ relationship that links diseases and anatomical sites in SNOMED CT. In this case, target ontologies of interest could be the SNOMED CT anatomy branch or the Foundational Model of Anatomy (FMA) [31], although mappings between SNOMED CT and FMA are not yet available in the UMLS Metathesaurus.

Our approach also acts as an alternative in scenarios where well-established subset extraction techniques might suffer from difficulties. This is the case for term frequency analysis when a local corpus related to the domain is not available or is not representative enough, and for ontology modularization techniques when they are not capable of working with multiple ontologies or there is uncertainty regarding which concepts to use as an input signature.

The analysis of our results revealed several limitations in the subset extraction process and opened new research questions to be explored in future work, as follows:

1. Improving the mapping technique beyond mere CUI matching by using the hierarchies of the ontologies to infer concepts that are now not taken into account, to minimize the impact of different granularity or term usage between ontologies.
2. Using other drug-indications linking components, such as the ones identified by Névéol and Lu [30].

3. Exploring our strategy in open biomedical ontology repositories such as BioPortal [32]. Open biomedical ontology repositories constitute a new opportunity to reuse domain-specific subsets submitted by users.
4. Quantitatively analyzing the performance gain (e.g., speed and memory usage) when using ontology subsets instead of full ontologies as thesauri in annotation applications.

Finally, this study adds to the existing studies suggesting that there is no universal way to extract subsets from ontologies and that the task of subset extraction should be strongly guided by each particular domain and application [14]. Furthermore, not only is the adequate technique essential to optimize the efficacy of the subsets, but also the data used for input and validation. More corpora of discharge summaries available to the health informatics community, which are still scarce, would also be especially welcome to generalize the results of this study to other domains and applications.

Acknowledgments

The authors thank everyone who contributed to this study in form of discussions, advice, reviews, and materials. We especially thank Nigam Shah, Natasha Noy, Timothy Redmond, Matthew Horridge, and Manuel Salvadores at the Stanford Center for Biomedical Informatics Research; Jesús Bermúdez, and Alfredo Goñi at the University of the Basque Country UPV/EHU; Rachel Richesson, Stefan Schulz, Martin Boeker, and Ronald Cornet during the 2012 International Conference on Biomedical Ontology; and Olivier Bodenreider at the National Library of Medicine. The authors also thank the U.S. National Institutes of Health for providing MEDLINE and the UMLS; the Massachusetts Institute of Technology, Philips Medical Systems, Philips Research North America, and Beth Israel Deaconess Medical Center for providing the MIMIC-II Clinical Database; and Partners HealthCare, and Informatics for Integrating Biology and the Bedside for providing the 2008 NLP Obesity Challenge corpus.

This work was supported by the Ministerio de Economía y Competitividad of the Spanish government (Grant TIN2010-21387-C02-01) and by the U.S. National Center for Biomedical Ontology (Grant HG004028 from the U.S. National Institutes of Health).

References

1. Pathak J, Johnson TM, Chute CG. Survey of modular ontology techniques and their applications in the biomedical domain. Integr Comput Aided Eng. 2009; 16(3):225–42. [PubMed: 21686030]
2. International Health Terminology Standards Organisation. [accessed April 2013] 2013. http://www.ihtsdo.org/
3. Baader, F.; Calvanese, D.; McGuinness, DL.; Nardi, D.; Patel-Schneider, PF., editors. The description logic handbook. Cambridge (UK): Cambridge University Press; 2003.
4. Stuckenschmidt, H.; Parent, C.; Spaccapietra, S., editors. Modular ontologies: concepts, theories and techniques for knowledge modularization. Springer; 2009.
5. Dai, M.; Shah, NH.; Xuan, W.; Musen, MA.; Watson, SJ.; Athey, BD., et al. An efficient solution for mapping free text to ontology terms. AMIA summit on translational bioinformatics; San Francisco, CA. 2008.
6. Jonquet, C.; Shah, N.; Youn, C.; Callendar, C.; Storey, M.; Musen, M. NCBO annotator: semantic annotation of biomedical data. Poster presented at the 8th International Semantic Web Conference (ISWC); 2009.
7. Euzenat, J.; Shvaiko, P. Ontology matching. Springer; 2007.
8. Navigli R. Word sense disambiguation: a survey. ACM Comput Surv. 2009; 41(2):1–69.
9. Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. J Am Med Infor Assoc. 2010; 17(3):229–36. http://dx.doi.org/10.1136/jamia.2009.002733.
10. Schuyler PL, Hole WT, Tuttle M, Sherertz D. The UMLS metathesaurus: representing different views of biomedical concepts. Bull Med Libr Assoc. 1993; 81(2):217–22. [PubMed: 8472007]
11. Noy, NF.; Musen, MA. Specifying ontology views by traversal. Proceedings of the 2004 international semantic web conference, lecture notes in computer science; Springer; p. 713-25.http://dx.doi.org/10.1007/978-3-540-30475-3_49

12. Seidenberg, J.; Rector, A. Web ontology segmentation: analysis, classification and use. Proceedings of the 15th international World Wide Web conference (WWW); 2006. p. 13-22.
13. Doran, P.; Tamma, V.; Iannone, L. Ontology module extraction for ontology reuse: an ontology engineering perspective. Proceedings of the sixteenth ACM conference on information and knowledge management; 2007. p. 61-70.
14. D'Aquin, M.; Schlicht, A.; Stuckenschmidt, H.; Sabou, M. Ontology modularization for knowledge selection: experiments and evaluations. Proceedings of the 18th international conference on Database and Expert Systems Applications (DEXA); 2007. p. 874-83.
15. Cuenca Grau B, Horrocks I, Kazakov Y. Modular reuse of ontologies: theory and practice. J Artif Intell Res. 2008; 31(1):273–318.
16. Vescovo, CD.; Parsia, B.; Sattler, U.; Schneider, T. The modular structure of an ontology: atomic decomposition. Proceedings of the 22nd international join conference on artificial intelligence; 2011. p. 2232-7.
17. López-García P, Boeker M, Illarramendi A, Schulz S. Usability-driven pruning of large ontologies: the case of SNOMED CT. J Am Med Inform Assoc. 2012; 19:e102–9. http://dx.doi.org/10.1136/amiajnl-2011-000503. [PubMed: 22268217]
18. Patrick J, Wang Y, Budd P, Rector A, Brandt S, Rogers J, et al. Developing SNOMED CT subsets from clinical notes for intensive care service. Health Care Inform Rev Online. 2008; 12(3):25–30.
19. Xu, R.; Musen, MA.; Shah, NH. A comprehensive analysis of five million UMLS metathesaurus terms using eighteen million MEDLINE citations. Proceedings of the AMIA 2010 annual symposium; 2010. p. 907-11.
20. Fung KW, McDonald C, Srinivasan S. The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institutions. J Am Med Inform Assoc. 2010; 17(6):675–80. http://dx.doi.org/10.1136/jamia.2010.007047. [PubMed: 20962130]
21. Saeed M, Villarroel M, Reisner A, Clifford G, Lehman L, Moody G, et al. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. Crit Care Med. 2011; 39(5):952–60. http://dx.doi.org/10.1097/CCM.0b013e31820a92c6. [PubMed: 21283005]
22. LePendu P, Iyer SV, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, et al. Pharmacovigilance using clinical notes. Clin Pharmacol Ther. 2013; 93(6):547–55. http://dx.doi.org/10.1038/clpt.2013.47. [PubMed: 23571773]
23. Xu R, Wang Q. Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing. BMC Bioinform. 2013; 14(1) http://dx.doi.org/10.1186/1471-2105-14-181 [Epub ahead of print].
24. Liu S, Ma W, Moore R, Ganesan V, Nelson SJ. Rxnorm: prescription for electronic drug information exchange. IT Prof. 2005; 7(5):17–23.
25. Burton, MM.; Simonaitis, L.; Schadow, G. Medication and indication linkage: a practical therapy for the problem list?. Proceedings of the 2008 AMIA annual symposium; 2008. p. 86-90.
26. U.S. Department of Affairs, VHA. [accessed April 2013] National Drug File–Reference (NDF-RT) terminology documentation. 2013. http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT%20Documentation.pdf
27. Brown SH, Elkin PL, Rosenbloom ST, Husser C, Bauer BA, Lincoln MJ, et al. VA national drug file reference terminology: a cross-institutional content coverage study. Stud Health Technol Inform. 2004; 107:477–81. [PubMed: 15360858]
28. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004; 32:D270–627.
29. Uzuner Ö. Recognizing obesity and comorbidities in sparse data. J Am Med Inform Assoc. 2009; 16:561–70. http://dx.doi.org/10.1197/jamia.M3115. [PubMed: 19390096]
30. Névéol, A.; Lu, Z. Automatic integration of drug indications from multiple health resources. Proceedings of the 1st ACM international health informatics symposium; ACM. 2010. p. 666-73.
31. Rosse C, Mejino JL Jr. A reference ontology for biomedical informatics: the foundational model of anatomy. J Biomed Inform. 2003; 36(6):478–500. [PubMed: 14759820]

32. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 2009; 37:W170–3. http://dx.doi.org/10.1093/nar/gkp440. [PubMed: 19483092]

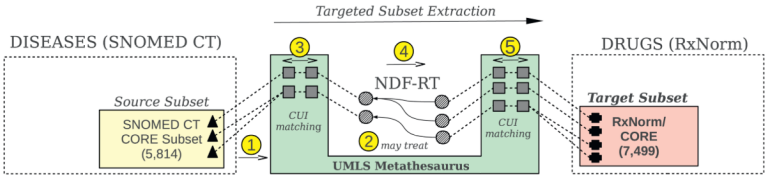

Fig. 1.

Steps followed to obtain drugs in RxNorm related to diseases in the SNOMED CT CORE subset, using NDF-RT and the UMLS Metathesaurus. We term the target subset RxNorm/CORE.

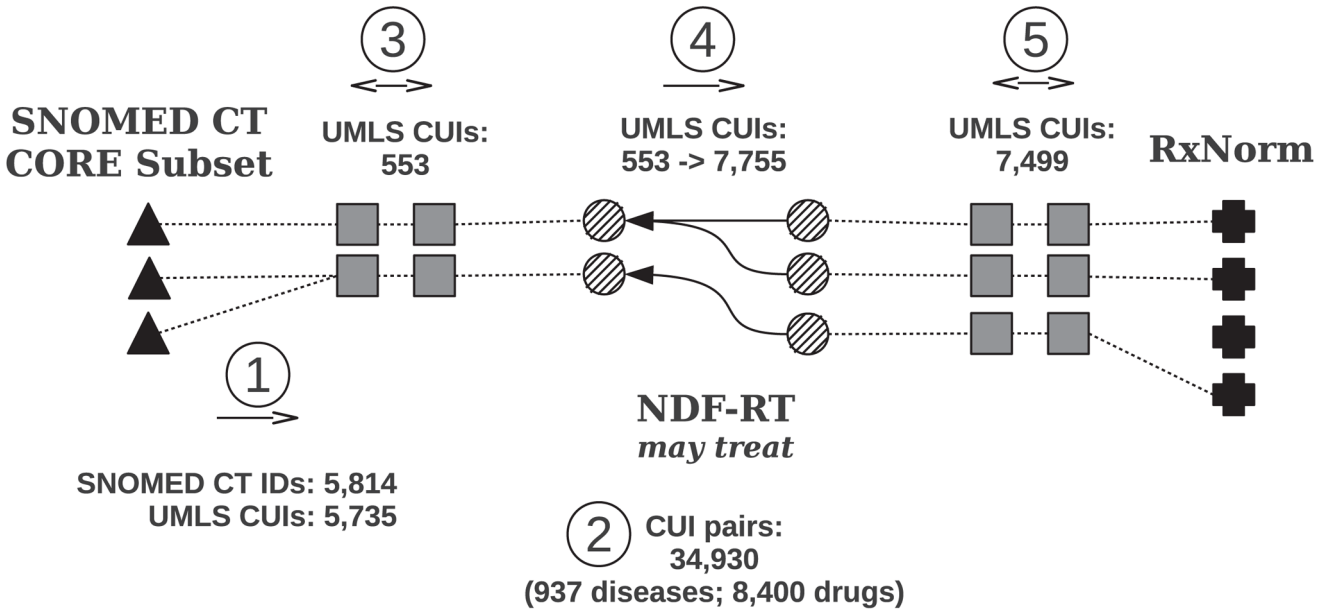

Fig. 2.
UMLS CUIs involved when extracting drugs in RxNorm related to diseases in the CORE problem list subset of SNOMED CT.

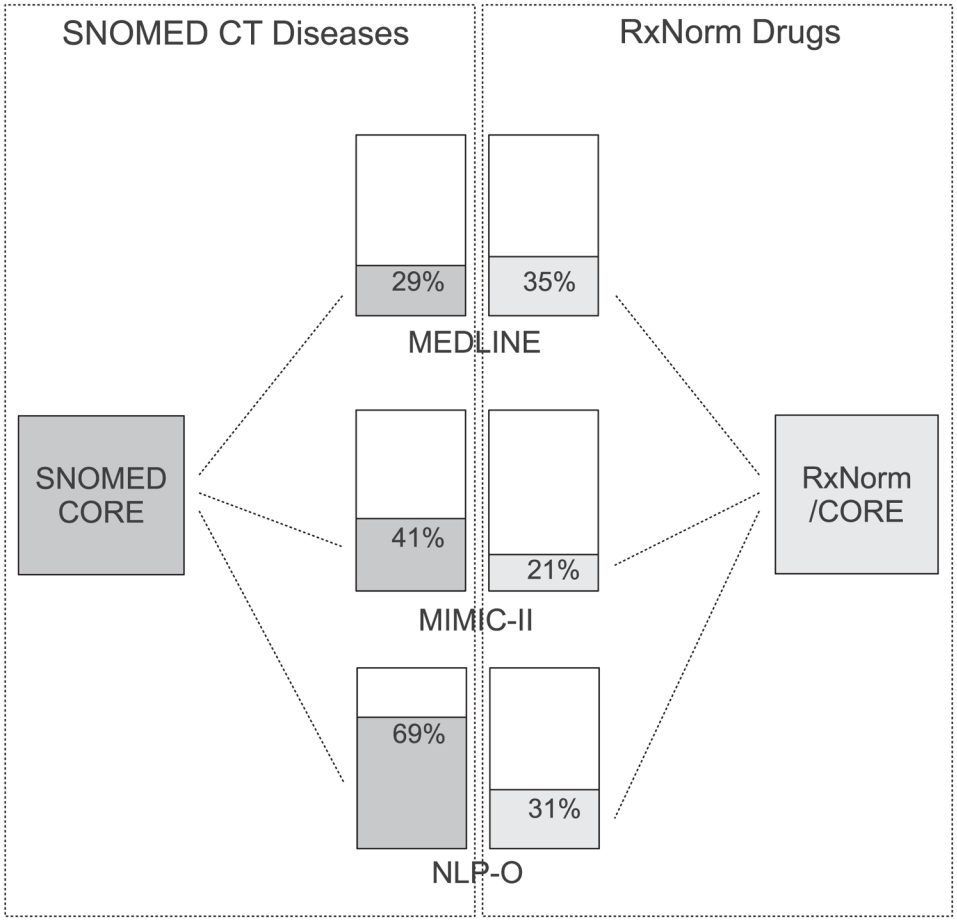

Fig. 3.
Recall comparison between the SNOMED CT CORE subset and RxNorm/CORE with respect to their reference ontologies.

Table 1

The most frequent concepts in the CORE problem list subset of SNOMED CT.

Concept ID	Preferred term	Usage (%)
38341003	Hypertensive disorder, systemic arterial	3.09
55822004	Hyperlipidemia	1.90
35489007	Depressive disorder	1.52
268565007	Adult health examination	1.37
235595009	Gastroesophageal reflux disease	1.23

Table 2

Distribution of concepts in the RxNorm/CORE subset, according to RxNorm term types.

Term type	CUIs (%)
Semantic Clinical Drug (SCD)	74.80
Ingredient (IN)	13.73
Precise Ingredient (PIN)	6.50
Semantic Branded Drug (SBD)	4.51
Designated Synonym (SY)	<1
Semantic Clinical Drug and Form (SCDF)	<1
Generic Drug Delivery Device (GPCK)	<1
Fully-specified drug Brand Name that cannot be prescribed (BN)	<1
Name for a Multi-Ingredient (MIN)	<1

Table 3

Extract of recommended treatments for edema, as captured by the ‘may treat’ relationship in NDF-RT.

Term	Term type	CUI
Torsemide	IN	C0076840
Torsemide 10 MG Oral Tablet	SCD	C0690835
Torsemide 10 MG/ML Injectable Solution	SCD	C0499011
Torsemide 100 MG Oral Tablet	SCD	C0690836
Torsemide 20 MG Oral Tablet	SCD	C0690837
Torsemide 5 MG Oral Tablet	SCD	C0690838
Triamterene	IN	C0040869
Triamterene 100 MG Oral Capsule	SCD	C0690636
Triamterene 50 MG Oral Capsule	SCD	C0690637
Trichlormethiazide	IN	C0040899
Trichlormethiazide 2 MG Oral Tablet	SCD	C0703753
Trichlormethiazide 4 MG Oral Tablet	SCD	C0690644

Table 4

Performance of RxNorm/CORE and CORE to annotate the selected corpora. The size and performance were measured with respect to their reference ontologies (RxNorm for RxNorm/CORE and SNOMED CT clinical findings/diseases branches for CORE).

Subset (reference ontology)	MEDLINE (%)	MIMIC-II (%)	NLP-O (%)
RxNorm/CORE (RxNorm)			
Relative size: 3.87%			
Precision	14.83	11.79	5.67
Recall	34.79	20.53	31.32
CORE (SNOMED CT diseases)			
Relative size: 9.10%			
Precision	31.06	25.95	9.22
Recall	28.60	40.84	69.34
CORE (SNOMED CT clinical findings)			
Relative size: 6.01%			
Precision	37.62	31.53	12.35
Recall	24.42	32.51	53.03

Table 5

The number of documents in each of the selected corpora for annotation.

Corpus	Documents
MEDLINE	206,484
MIMIC-II	26,657
NLP-O	611

Table 6

The number of concepts in each domain ontology and subset.

Domain	Ontology or subset	Concepts
Drugs	RxNorm/CORE	7499
	RxNorm	193,737
Diseases	CORE	5814
	SNOMED CT diseases	63,884
	SNOMED CT clinical findings	96,783

Table 7

Relevant concepts for each domain ontology and subset in the selected corpora. A concept was considered relevant if it appeared in at least one document of the corpus.

Domain	Ontology or subset	MEDLINE	MIMIC-II	NLP-O
Drugs	RxNorm/CORE	1112	884	425
	RxNorm	3196	4306	1357
Diseases	CORE diseases	1806	1509	536
	CORE clinical findings	2187	1833	718
	SNOMED CT diseases	6315	3695	773
	SNOMED CT clinical findings	8957	5638	1354