RESEARCH   Open Access

Improved characterisation of clinical text through ontology-based vocabulary expansion

Luke T. Slater[1,2]* , William Bradlow[1,2], Simon Ball[1,2], Robert Hoehndorf[3] and Georgios V Gkoutos[1,2,4,5,6,7]

Abstract

Background: Biomedical ontologies contain a wealth of metadata that constitutes a fundamental infrastructural resource for text mining. For several reasons, redundancies exist in the ontology ecosystem, which lead to the same entities being described by several concepts in the same or similar contexts across several ontologies. While these concepts describe the same entities, they contain different sets of complementary metadata. Linking these definitions to make use of their combined metadata could lead to improved performance in ontology-based information retrieval, extraction, and analysis tasks.

Results: We develop and present an algorithm that expands the set of labels associated with an ontology class using a combination of strict lexical matching and cross-ontology reasoner-enabled equivalency queries. Across all disease terms in the Disease Ontology, the approach found **51,362** additional labels, more than tripling the number defined by the ontology itself. Manual validation by a clinical expert on a random sampling of expanded synonyms over the Human Phenotype Ontology yielded a precision of **0.912**. Furthermore, we found that annotating patient visits in MIMIC-III with an extended set of Disease Ontology labels led to semantic similarity score derived from those labels being a significantly better predictor of matching first diagnosis, with a mean average precision of **0.88** for the unexpanded set of annotations, and **0.913** for the expanded set.

Conclusions: Inter-ontology synonym expansion can lead to a vast increase in the scale of vocabulary available for text mining applications. While the accuracy of the extended vocabulary is not perfect, it nevertheless led to a significantly improved ontology-based characterisation of patients from text in one setting. Furthermore, where run-on error is not acceptable, the technique can be used to provide candidate synonyms which can be checked by a domain expert.

Keywords: Text mining, Ontology, Vocabulary expansion, Semantic similarity

*Correspondence: l.slater.1@bham.ac.uk
[1]Institute of Cancer and Genomic Sciences, College of Medical and Dental Sciences, University of Birmingham, B15 2TT Birmingham, UK
[2]University Hospitals Birmingham NHS Foundation Trust, University of Birmingham, B15 2TT Birmingham, UK
Full list of author information is available at the end of the article

Background

Metadata are a fundamental feature of biomedical ontologies, describing a wealth of natural language information in the form of labels and descriptions [1]. Ontologies formalized in the Web Ontology Language (OWL) [2] implement metadata in the form of annotation properties, and these can be used to describe multiple natural language labels for a single term, forming a collection of words and phrases that humans use to signify the concept. Open Biomedical Ontologies (OBO) [3] and the Information Artifact Ontology (IAO) [4] define a series of conventional annotation properties that can be used for the expression of labels and synonyms. These features are widely used: an investigation of ontologies in BioPortal found that 90% of classes had a label associated with them [5]. For example, as of 2017 the Human Phenotype Ontology (HP) [6] contained 14,328 synonyms for 11,813 classes [7]. The labels associated with ontology terms constitute a controlled domain vocabulary [1].

The domain vocabulary makes ontologies a valuable resource for text mining, particularly in information retrieval and extraction tasks [8]. The natural language labels associated with ontology classes can be used to identify where a class is mentioned in text. Furthermore, association of entities described in text with ontologies enables their integration with other datasets annotated by the same ontologies, as well as caters to the application of ontology-based analysis techniques such as semantic similarity [9, 10], semantic data mining [11], machine learning [12], or clustering [13].

However, due to limitations on resources for expert curation of ontologies and the sheer scale of their contents, the labels obtainable from single ontologies are not exhaustive. Combined with the tendency for alternative presentation of semantically equivalent concepts in biomedical text [14], ontology labels are not always a good fit for text corpora that mention the entities described by ontology concepts [15]. By expanding the set of synonyms in an ontology, particularly with synonyms that provide a better fit for text corpora, the performance of natural language processing tasks that depend on them may be improved.

This potential is reflected by previous work in the field. One approach that used analysis of existing synonyms across ontology hierarchy to determine new synonyms reported an increase in performance for the task of retrieving articles from a literature repository [16]. Another rule-based synonym expansion approach to extending the Gene Ontology showed improved performance in concept recognition tasks [17]. A combined machine-learning and rule-based approach to learning new HP synonyms from manually annotated PubMed abstracts improved performance of an annotation task over a gold standard text corpus [18]. These methods combine label components with label components from upper level classes to generate additional candidate synonyms, and search text corpora to limit those to true synonyms.

Ontology-based annotation software such as OBO Annotator [19], ConceptMapper [20], and the NCBO Annotator [21] contain routines to consider rule-based morphological and positional transformations of terms to increase concept recognition recall. Parameters that control the use of these features have a strong influence on annotation performance [22]. Previous work has also investigated synonym acquisition and derivation for the purposes of improving the performance of lexical ontology matching and alignment tasks [23]. Outside of automated synonym generation, organised efforts have been made to manually extend an ontology's synonyms for a particular purpose. For example, HP was expanded with layperson synonyms to enable its use in applications that interact directly with patients [24].

However, no work to our knowledge has considered linking different ontology classes for the purposes of vocabulary expansion. Many biomedical entities are described by several classes in equivalent or similar contexts across several ontologies. For example, terms describing hypertension exist in many ontologies and medical terminologies. The *hypertension* (HP:0000822) term describes the condition in the context of a phenotype, while *hypertension* (DOID:10763) from the Disease Ontology (DO) [25] describes it in the context of a disease (although the difference between a disease and a phenotype is disputed). Specific-disease or application ontologies also extend upon definitions provided by general domain ontologies. For example, the Hypertension Ontology (HTN) [26] extends the HP and DO hypertension classes, adding additional information including labels. Furthermore, the subtle distinctions between concepts that biomedical ontologies capture, including phenotype versus disease, do not necessarily influence many of the commonly applied text mining tasks, because these contexts share the same labels.

We hypothesise that because ontologies are constructed with different focuses, ontologies that define concepts describing the same real-world entities will contain different, but valid, synonyms for a particular context. These focuses consist in contexts, domain experts, and source material. By considering all of these terms, we can construct extended vocabularies that may improve the power of ontology-based text mining tasks.

In this paper, we describe and implement a synonym expansion approach that combines lexical matching and semantic equivalency to obtain new synonyms for biomedical concepts. The synonym expansion algorithm derives additional synonyms for a class by matching it with classes from other ontologies, making use of the

AberOWL ontology reasoning framework [27]. We use the approach to extend several ontology vocabularies, and evaluate them both manually, and in an ontology-based patient characterisation task.

Results

The synonym expansion algorithm is available as part of the Komenti text mining framework, which is available under an open source licence at https://github.com/reality/komenti, while the files used for validation are available at https://github.com/reality/synonym_expansion_validation.

Algorithm

The synonym expansion algorithm, including the two matching methods and steps to prune candidate synonyms, is described below. The process is performed for every input class given (in this context, 'every ontology' is any of the ontologies that are included in AberOWL). An input class is any ontology class for which we want to obtain additional synonyms.

1. Extract the labels and synonyms of any classes in any ontology with a label or synonym that exactly matches the first label of the input class.
2. Run an equivalency query against every ontology using the Internationalised Resource Identifier (IRI) of the input class, extracting labels and synonyms for any classes returned.
3. Of the candidate synonyms produced by the first two steps, discard any that were:

   - Defined in ontologies that were found to produce incorrect synonyms.
   - Have the form of a term identifier.
   - Contain the input class label as a substring.

The algorithm uses two different methods for identifying matching classes, specified in steps one and two above. Strict lexical matching is used to identify otherwise unlinked terms that contain a label which is the same as the first label of the input class. Only the first label for the input class is used, because we found that the additional labels and synonyms were more likely to match classes which had different meanings, and led to more incorrect candidate synonyms. Mapping terms across ontologies via shared labels or metadata is a well established technique used in ontology alignment [28].

Equivalency queries are used to obtain additional candidate synonyms from classes that are equivalent to the input class, but do not share the same first label. In OWL ontologies, classes are uniquely identified by their IRIs, and classes that share the same IRI are automatically considered equivalent by a reasoner. This can be used to match classes in the case that another use of the same class

is not expressed with the same first label in another ontology, occurring due to ontologies becoming out of sync, or intentional omission of annotation properties in a referencing class. In addition, equivalencies between different classes can be directly asserted via axioms in an ontology, or can appear as the result of a logical inference. Since the classes are semantically equivalent, we can use the metadata, including labels, of the other class to refer to the original. To retrieve equivalent classes, the AberOWL API runs an equivalency query against each ontology in the repository, which uses the description logic reasoner to obtain a list of matching classes, which are used to contribute additional synonyms.

After the main matching stage, the set of labels is pruned down to remove incorrect values. Some ontologies include term identifiers as labels which cannot be exploited by text-mining applications. Therefore, candidate synonyms that contained a colon or underscore were removed. The algorithm also removes labels sourced from GO-PLUS [29], MONDO [30], CCONT [31], and phenX [32], because we found these ontologies consistently produced incorrect synonyms. Incorrect synonyms could be contained in these ontologies due to human error, or in the case of large meta-ontologies such as MONDO, as a result of algorithmic error in asserting equivalencies between phenotypes across species. We also removed labels that include the input label as a substring, as these add no value to concept recognition systems (as the smaller string would match, making the longer string redundant).

Ontology expansion

We applied the vocabulary expansion algorithm to all 9,908 subclasses of *disease* (DOID:4) in the Disease Ontology (DO). DO itself asserts 24,878 labels and synonyms for these classes. The expanded DO vocabulary contained 76,240 labels and synonyms. We also applied the algorithm to the 14,406 non-obsolete subclasses of *Phenotypic abnormality* (HP:0000118) in HP. HP itself asserts 29,805 labels and synonyms. The number of labels and synonyms following expansion was 54,765. Therefore, the algorithm found 24,960 additional synonyms for terms in HP.

For the DO term *hypertension* (DOID:10763), 28 labels and synonyms were found. 3 of these were from DO itself. The first two steps of the algorithm, which obtains candidate synonyms, found 70 synonyms not including the word 'hypertension' itself. Of these, 56 were obtained via lexical matching, and 14 by equivalency query. The sources of these synonyms are summarised in Table 1. After making the list unique, there were 28 labels and synonyms. Therefore, the algorithm found 25 new synonyms, that were not asserted in the original DO term.

In this example, there were no synonyms uniquely found via equivalency. However, if we use *bradycardia* as the input class, we can identify two new synonyms from

Table 1 Source of the 70 non-unique synonyms found for the term *hypertension* (`DOID:10763`) per-ontology

Ontology	Source	Number of Synonyms
GWAS_EFO_SKOS	Lexical	16
MESH	Lexical	4
HTN [26]	Lexical, Equivalency	3 (2)
CRISP	Lexical	1
CCTOO [33]	Lexical	6
ONTONEO	Lexical, Equivalency	4 (3)
NCIT [34]	Lexical	6
COSTART	Lexical	7
BAO	Lexical, Equivalency	2
CSSO	Lexical	2
ODAE	Lexical, Equivalency	2
DO	–	3
DTO	Equivalency	2
Total	–	70
Total Unique	–	28

Of these synonyms, 28 were unique. The source column describes which class matching methods were used to match classes that contributed synonyms from the external ontology. Lexical refers to when classes were found through a matching first label, and equivalency through a semantic equivalency query. Bracketed numbers, where given, are the labels found by the equivalency method only

PhenomeNET [35], *bradyrhythmia* and *reduced heart rate*, which were not otherwise obtained via lexical matching. This is because PhenomeNET establishes a semantic equivalence between *decreased heart rate* (MP:0005333) and *bradycardia*, which does not share its first label with the HP class.

Manual validation

To evaluate the correctness of synonyms in the expansion of HP, a clinical expert manually evaluated 866 novel synonyms found for 500 randomly selected terms. Table 2 summarises the results, which show a precision of 0.912. 195 terms were marked as ambiguous, in the case that the synonyms were in a foreign language or the clinician did not have enough expertise of the term to determine whether the synonym was correct. Of these, the vast majority (161) were non-English labels, while the remaining 32 were English language synonyms the clinician could not judge.

Annotation

To initially evaluate whether the extended vocabularies could lead to more annotations of biomedical text, which could lead to greater performance at information retrieval and extraction tasks, we annotated the text associated with 1,000 randomly sampled MIMIC-III patients. We built a vocabulary using all non-obsolete subclasses of *Abnormality of the cardiovascular system* (HP:0011025), and compared the number of annotations before and after vocabulary expansion using our method. HP asserts 2,205 labels and synonyms for these classes, while the expanded set of labels numbers 5,336. The results are summarised in Table 3.

Patient characterisation

While the annotation task showed that our method can lead to more annotations, this does not necessarily mean that those annotations were correct or informative. Indeed, the manual validation indicates that there is some level of error associated with the process. To identify whether the annotations were informative and useful, we evaluated how the increased number of annotations affected performance on a downstream task.

In particular, we evaluated whether the additional ontology annotations yielded by the vocabulary expansion process led to better performance in using semantic similarity calculated from those annotations to predict shared primary diagnosis within the MIMIC-III dataset [36]. We annotated a sample of 1,000 patient visits using classes from the Disease Ontology (DO) that contained cross-references to ICD-9, both before and after label expansion using the presented algorithm. We then used those annotations to calculate a measure of semantic similarity between the patient visits, and evaluated the rankings with respect to whether highly ranked patient visits shared the primary diagnosis ICD-9 code (which each patient visit is annotated with in MIMIC), including those we did not find through DO cross-references (and were therefore not annotated).

The semantic similarity approach allows us to match patients who share a primary diagnosis even if they are not annotated directly with that disease (in this case, if we did not have an ICD-9 mapping for that disease), under the assumption patients who share the same diagnosis will be more similar on the basis of auxiliary symptoms associated with the disease they share. If we can more effectively annotate patients with the conditions that we do know

Table 2 Metrics for clinical expert validation of 866 generated synonyms for 500 terms

Terms	Total Synonyms	TP	FP	Non-English	Uncertain	Precision
500	866	614	59	161	32	0.912 (0.709)

Synonyms already included in HP were not included in the validation. Synonyms were either marked correct, incorrect, non-English, or uncertain. Uncertain was chosen if the validator did not have enough expertise to confidently judge the synonym. Precision is calculated with TP and FP columns of the table, while the figure in parentheses is calculated using the sum of the FP, Non-English, Uncertain columns as false positives, to illustrate the worst case scenario, where every unknown synonym is actually incorrect

Table 3 Amount of labels for *Abnormality of the cardiovascular system* (HP:0011025) before and after synonym expansion, and the amount of annotations made of text associated with 1,000 MIMIC patients with these vocabularies

Vocabulary	Labels	MIMIC-III Annotations
HP Labels	2,205	1,104
Expanded HP Labels	5,336	1,447

about (present in our annotation vocabulary), then we should be able to rank them together in a way that is better predictive of a shared primary diagnosis.

We used the mean reciprocal rank and the mean average precision to measure how well semantic similarity rankings predicted matching first diagnoses. The results of the ranking task are shown in Table 4, with the expanded vocabulary leading to an increased performance in both cases. To determine whether the result was significantly different, we used the Wilcoxon rank-sum test to compare the ranks of patient similarity pairs with matching first diagnoses, yielding a *p*-value of 0.0007063.

Discussion

The results clearly demonstrate that for two biomedical ontologies, our approach vastly increases the amount of labels and synonyms available for their terms. Using hypertension as an example, we demonstrated that a range of different ontologies contribute additional synonyms, leading to 25 new unique labels for the term. By leveraging these we can effectively enrich vocabularies for terms.

While we only manually validated a small subset of terms from HP, this indicated a fairly high precision for candidate terms. Through analysis of the false positives, we found that many of them were caused by errors in the ontologies that the synonyms were sourced from. For example, several synonyms for *motor aphasia* (HP:0002427) were marked as incorrect since they refer to dysphasia, including "Broca Dysphasia." Aphasia and dysphasia are different conditions. The first refers to a partial loss of language, and the latter to a full loss of language. All of these incorrect synonyms were sourced from *Aphasia, Broca* (MESH:D001039) in MESH.

Table 4 Comparison of the annotations of texts for 1,000 randomly sampled MIMIC-III patient visits before and after expansion, and their associated performance with respect to how predictive semantic similarity scores calculated from the annotations were of shared first diagnosis

Investigation	Annotations	MAP	MRR
Unexpanded	1,380,216	0.88	0.947
Expanded	2,088,765	0.913	0.986

Though this is not reflected in the results, we also found during the development of the algorithm that certain ontologies produced consistently incorrect synonyms. Several of these ontologies are meta-ontologies, automatically constructed from several ontologies using alignment and integration methods, and it is possible that errors in that process were the cause of the incorrect synonyms. Certain annotation properties were also incorrectly detailed by the AberOWL API as being labels, such as *europe pmc* and *kegg compound*. Candidate synonyms defined by problematic ontologies or matching the list of annotation properties are automatically removed. Expansion of the list of ontologies discluded from the sources for labels might further improve the precision of the algorithm, but may potentially come at the cost of correct synonyms.

Furthermore, the manual validation revealed that many of the returned synonyms were in non-English languages. While OWL ontologies do allow for parameters that distinguish which language the property is in, AberOWL does not index them. Therefore, it is not currently possible to distinguish between English and non-English synonyms. These items were marked as ambiguous, and not counted in the overall precision. This could also be controlled partially by discluding additional ontologies from results. For example, WHOFRE is a non-ontology mapping of French vocabulary to UMLS. For any uses where a reduced vocabulary accuracy is not acceptable, the algorithm should be used as a candidate label generator, to be checked by a domain expert before further use.

We also demonstrated that our expansion of the HP vocabulary increases the amount of phenotype annotations produced for MIMIC-III patient visit text records. While we did not directly validate the correctness of these annotations, by necessity a time-consuming task, we explored whether the additional synonyms would improve performance in a downstream task in our final evaluation. This experimental evaluation showed a clear and significant increase for a patient stratification task over MIMIC-III, identifying shared first diagnosis via semantic similarity score derived from ontology annotations. This indicates that for certain tasks, our approach can increase the quality of entity characterisations gained by information extraction, and in turn the power of ontology-based analyses, even without manual validation of the produced labels.

Limitations and future work

The most important potential limitation of the algorithm itself is that it violates the notion that the IRI of a concept uniquely identifies it, rather than its name. This is due to the fact that OWL ontologies do not follow the unique name assumption. False positives, in theory, could be generated by a lexical match on a homonym, which

Slater *et al. Journal of Biomedical Semantics*   (2021) 12:7

Page 6 of 9

then has different synonyms itself. We believe, however, that this effect should be limited in the case of a highly specific biomedical language. Furthermore, any such error would be most likely be mitigated by the dataset context limitation. For example, synonyms derived from different contexts, incorrectly associated with a medical concept, are unlikely to be present within clinical letters.

False synonyms could also be removed on the basis of a corpus search. For example, if a candidate synonym never, or, at least, rarely, appears in the same document as another label, used for this term across a literature corpus, it is possible that it refers to a different concept from a disjoint context. This could also be performed by analysing the metadata of text corpora. For example, if two terms are never, or, at least, rarely, associated with literature from the same journals, the same field, or the same content tags, it is possible they have different meanings. In a further study, we would investigate whether synonymy can be identified using word embeddings.

While equivalency returns fewer synonyms, and not necessarily many that are not also obtained by lexical matching, they can also be treated with a higher level of confidence. For this reason, using only this method could be considered as a parameter in the case that a higher accuracy is required.

Conclusions

We have demonstrated that an inter-ontology approach to vocabulary expansion is a powerful method for adding informative labels and synonyms to terms used in text mining. These synonyms are found with a fairly high precision, and led to a greater rate of document retrieval in clinical and literature settings. Most importantly, we have shown that the approach improves the power of an ontology-based characterisation and analysis of patients via clinical text.

Methods

All files described in the validation (excluding the MIMIC-III data files), along with the commands necessary to repeat the experiments are available at https://github.com/reality/synonym_expansion_validation/.

Algorithm

We implemented the algorithm as a module in the Komenti semantic text mining framework using the Groovy programming language [37]. It makes use of the AberOWL API [27] for label matching and semantic queries, documented at http://www.aber-owl.net/docs/.

OWL ontologies use a number of conventional annotation properties to define labels and synonyms. These span a range of confidence and degree of synonymy. In this paper, we consider frequently used annotation properties, summarised in Table 5. These are the annotation

properties consolidated into the 'synonym' property by the AberOWL API. Another oboInOwl synonym, *hasRelatedSynonym* is excluded, because the labels provided by these synonyms are too imprecise.

Manual validation

To evaluate the performance of the algorithm, we randomly selected 500 classes from the expanded version of HP for manual validation. Synonyms already asserted by HP were removed from the set, because they were already assumed to be correct, and would not contribute to measuring the performance of the synonym expansion algorithm. A clinical expert (WB) marked each synonym as correct, incorrect, or ambiguous. The expert was asked to answer correctly or incorrectly on the basis: "if a patient has *synonym*, would it also be true that they have *original label*?" Entries were marked as ambiguous if the synonym was in a different language, or the validator otherwise did not have enough knowledge of the phenotype to determine whether or not the synonym was correct.

Annotation

We used the Komenti semantic text mining framework, which implements Stanford CoreNLP's RegexNER [38] to annotate 1,000 randomly sampled entries from the NOTEEVENTS table in MIMIC-III (MIMIC) [39]. MIMIC is a freely available healthcare database, containing a variety of structured and unstructured information concerning around 60,000 admissions to critical care services [36]. We annotated the sample with all subclasses of *Abnormality of the cardiovascular system* (HP:0011025), comparing the number of annotations before and after synonym expansion. This investigation was performed on 17/01/2020.

Patient characterisation

We sampled 1,000 patient visits from the MIMIC-III (distinct from those used in the annotation experiment). We then concatenated all text records for each patient visit from the NOTEEVENTS table into one text file, and pre-processed the text to remove newlines, improve sentence delineation, and lemmatise words. We also retained the primary diagnosis, which was the first listed ICD-9 code in the DIAGNOSES_ICD table. These codes are produced by clinical coding specialists, by examining the texts associated with the visit.

We limited the classes considered for our annotation vocabulary to those which DO contained a database cross-reference to ICD-9, of which there were 2,118. This was to reduce noise from terms not represented in ICD-9. We obtained the unexpanded and expanded synonyms for these terms on 08/07/2020. Both sets of labels were also lemmatised (both lemmatised and unlemmatised forms were used for annotation).

Table 5 Summary of conventionally used annotation properties considered in this experiment

Annotation Property	Identifier	Definition
label	rdfs:label	"a human-readable version of a resource's name [40]."
altLabel	skos:core#altLabel	"An alternative lexical label for a resource [41]."
has_exact_synonym	hasExactSynonym	"An alias in which the alias exhibits true synonymy [42]."
has_narrow_synonym	hasNarrowSynonym	"An alias in which the alias is narrower than the primary class name. Example: pyrimidine-dimer repair by photolyase is a narrow synonym of photoreactive repair [42]."
has_broad_synonym	hasBroadSynonym	"An alias in which the alias is broader than the primary class name. Example: cell division is a broad synonym of cytokinesis [42]."
alternative term	IAO_0000118	"An alternative name for a class or property which means the same thing as the preferred name (semantically equivalent) [4]."

Definitions come from the description of the annotation properties in their respective top-level ontologies.

The Komenti semantic text-mining framework was used to annotate the text associated with each patient visit. As before, this made use of the CoreNLP RegexNER annotator [38]. Negated annotations were excluded using the komenti-negation algorithm [43]. We then used the set of terms associated with it to produce a semantic similarity matrix for patient visits, using the Resnik measure of pairwise similarity for each annotated term [10], normalised into a groupwise measure using the best match average method [9]. Information content was calculated using the probability of the term appearing as an annotation in the totality of the set of annotations [10]. The similarity matrix was computed using the Semantic Measures Library [44].

We evaluated the similarity matrix using mean reciprocal rank and mean average precision to measure performance in predicting shared primary patient diagnosis. A true case was considered to be whether a pair of patient visits had the same primary diagnosis (as per the MIMIC-III database). For mean average precision, we considered only the 10 most similar patients for each patient. The *p*-value was calculated using the built-in *wilcoxon.test* function of R version 3.4.4 [45].

Abbreviations
OWL: Web ontology language; OBO: Open biomedical ontologies; HP: Human phenotype ontology; DO: Disease ontology; NER: Named entity recognition; IRI: Internationalised resource identifier; MIMIC: MIMIC-III; IAO: Information artifact ontology

Authors' contributions
LTS conceived of the study, performed the experiments, implemented the software, and wrote the first draft the manuscript. RH conceived of the patient characterisation experiment. WB performed data validation and contributed to evaluation of results. All authors contributed in the analysis and interpretation of the results. RH and GVG contributed to the manuscript. GVG, RH, and SB supervised the project. All authors revised and approved the manuscript for submission.

Availability of data and materials
The implementation of the algorithm described is available under an open source licence at http://github.com/reality/komenti, while the experimental files (excluding the MIMIC-III data files) are available at https://github.com/reality/synonym_expansion_validation.

Declarations

Ethics approval and consent to participate
This work makes use of the MIMIC-III dataset, which was approved for construction, de-identification, and sharing by the BIDMC and MIT institutional review boards (IRBs). Further details on MIMIC-III ethics are available from its original publication (DOI:10.1038/sdata.2016.35). Further ethical approval was not required for this experiment, as it concerns a public dataset. Work was undertaken in accordance with the MIMIC-III guidelines.

Consent to publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Institute of Cancer and Genomic Sciences, College of Medical and Dental Sciences, University of Birmingham, B15 2TT Birmingham, UK. [2] University Hospitals Birmingham NHS Foundation Trust, University of Birmingham, B15 2TT Birmingham, UK. [3] Computational Bioscience Research Centre, KAUST, Thuwal, Saudi Arabia. [4] NIHR Experimental Cancer Medicine Centre, University of Birmingham, B15 2TT Birmingham, UK. [5] NIHR Surgical Reconstruction and Microbiology Research Centre, University of Birmingham, B15 2TT

Birmingham, UK. [6]NIHR Biomedical Research Centre, University of Birmingham, B15 2TT Birmingham, UK. [7]MRC Health Data Research (HDR), Birmingham, UK.

References

1. Hoehndorf R, Schofield PN, Gkoutos GV. The role of ontologies in biological and biomedical research: A functional perspective. Brief Bioinform. 2015;16(6):1069–80. https://doi.org/10.1093/bib/bbv011.
2. Grau BC, Horrocks I, Motik B, Parsia B, Patel-Schneider P, Sattler U. OWL 2: The next step for OWL. J Web Semant. 2008;6(4):309–22. https://doi.org/10.1016/j.websem.2008.05.001.
3. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone S-A, Scheuermann RH, Shah N, Whetzel PL, Lewis S. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25(11):1251–5. https://doi.org/10.1038/nbt1346.
4. Ceusters W. An information artifact ontology perspective on data collections and associated representational artifacts. MIE; 2012, p. 6872.
5. Quesada-Martínez M, Fernández-Breis JT, Stevens R. Lexical characterization and analysis of the BioPortal ontologies. In: Peek N, Marín Morales R, Peleg M, editors. Artificial Intelligence in Medicine. Berlin, Heidelberg: Springer; 2013. p. 206–15. https://doi.org/10.1007/978-3-642-38326-7_31.
6. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GCM, Brown DL, Brudno M, Campbell J, FitzPatrick DR, Eppig JT, Jackson AP, Freson K, Girdea M, Helbig I, Hurst JA, Jähn J, Jackson LG, Kelly AM, Ledbetter DH, Mansour S, Martin CL, Moss C, Mumford A, Ouwehand WH, Park S-M, Riggs ER, Scott RH, Sisodiya S, Vooren SV, Wapner RJ, Wilkie AOM, Wright CF, Vulto-van Silfhout AT, de Leeuw N, de Vries BBA, Washington NL, Smith CL, Westerfield M, Schofield P, Ruef BJ, Gkoutos GV, Haendel M, Smedley D, Lewis SE, Robinson PN. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. 2014;42(Database issue):966–74. https://doi.org/10.1093/nar/gkt1026.
7. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, Baynam G, Bello SM, Boerkoel CF, Boycott KM, Brudno M, Buske OJ, Chinnery PF, Cipriani V, Connell LE, Dawkins HJS, DeMare LE, Devereau AD, de Vries BBA, Firth HV, Freson K, Greene D, Hamosh A, Helbig I, Hum C, Jähn JA, James R, Krause R, F. Lauederkind SJ, Lochmüller H, Lyon GJ, Ogishima S, Olry A, Ouwehand WH, Pontikos N, Rath A, Schaefer F, Scott RH, Segal M, Sergouniotis PI, Sever R, Smith CL, Straub V, Thompson R, Turner C, Turro E, Veltman MWM, Vulliamy T, Yu J, von Ziegenweidt J, Zankl A, Züchner S, Zemojtel T, Jacobsen JOB, Groza T, Smedley D, Mungall CJ, Haendel M, Robinson PN. The Human Phenotype Ontology in 2017. Nucleic Acids Res. 2017;45(D1):865–76. https://doi.org/10.1093/nar/gkw1039.
8. Spasic I, Ananiadou S, McNaught J, Kumar A. Text mining and ontologies in biomedicine: Making sense of raw text. Brief Bioinform. 2005;6(3):239–51. https://doi.org/10.1093/bib/6.3.239.
9. Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007;23(10):1274–81. https://doi.org/10.1093/bioinformatics/btm087.
10. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007. 1995.
11. Dou D, Wang H, Liu H. Semantic data mining: A survey of ontology-based approaches. In: Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015); 2015. p. 244–51. https://doi.org/10.1109/ICOSC.2015.7050814.
12. Kulmanov M, Smaili FZ, Gao X, Hoehndorf R. Machine learning with biomedical ontologies. bioRxiv preprint. 20202020.05.07.082164.
13. Lin Y, Xiang Z, He Y. Towards a Semantic Web Application: Ontology-Driven Ortholog Clustering Analysis. In: Proceedings of ICBO 2011. 2011. p. 33–41.
14. Cohen KB, Palmer M, Hunter L. Nominalization and alternations in biomedical language. PLoS ONE. 2008;3(9):3158. https://doi.org/10.1371/journal.pone.0003158.
15. Brewster C, Alani H, Dasmahapatra S, Wilks Y. Data driven ontology evaluation. In: International Conference on Language Resources and Evaluation (30/05/04); 2004. https://www.aclweb.org/anthology/L04-1476/.
16. Taboada M, Rodriguez H, Gudivada RC, Martinez D. A new synonym-substitution method to enrich the human phenotype ontology. BMC Bioinformatics. 2017;18:446.
17. Funk CS, Cohen KB, Hunter LE, Verspoor KM. Gene Ontology synonym generation rules lead to increased performance in biomedical concept recognition. J Biomed Semant. 2016;7(1):52. https://doi.org/10.1186/s13326-016-0096-7.
18. Lobo M, Lamurias A, Couto FM. Identifying human phenotype terms by combining machine learning and validation rules. BioMed Res Int. 2017;2017:. https://doi.org/10.1155/2017/8565739.
19. Groza T, Kohler S, Doelken S, Collier N, Oellrich A, Smedley D, et al. Automatic concept recognition using the Human Phenotype Ontology reference and test suite corpora. Database. 2015;2015:bav005bav005.
20. Tanenblatt MA, Coden A, Sominsky IL. The ConceptMapper Approach to Named Entity Recognition. LREC: Citeseer; 2010, p. 54651.
21. Jonquet C, Shah N, Youn C, Callendar C, Storey M-A, Musen M. NCBO annotator: semantic annotation of biomedical data. Washington DC: International Semantic Web Conference, Poster and Demo session; 2009.
22. Funk C, Baumgartner W, Garcia B, Roeder C, Bada M, Cohen KB, Hunter LE, Verspoor K. Large-scale biomedical concept recognition: An evaluation of current automatic annotators and their parameters. BMC Bioinformatics. 2014;15:59. https://doi.org/10.1186/1471-2105-15-59.
23. Pesquita C, Faria D, Stroe C, Santos E, Cruz IF, Couto FM. What's in a 'nym'? Synonyms in Biomedical Ontology Matching. In: Alani H, Kagal L, Fokoue A, Groth P, Biemann C, Parreira JX, Aroyo L, Noy N, Welty C, Janowicz K, editors. The Semantic Web – ISWC 2013. Berlin, Heidelberg: Springer; 2013. p. 526–41. https://doi.org/10.1007/978-3-642-41335-3_33.
24. Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine J-P, Gargano M, Harris NL, Matentzoglu N, McMurry JA, Osumi-Sutherland D, Cipriani V, Balhoff JP, Conlin T, Blau H, Baynam G, Palmer R, Gratian D, Dawkins H, Segal M, Jansen AC, Muaz A, Chang WH, Bergerson J, Laulederkind SJF, Yüksel Z, Beltran S, Freeman AF, Sergouniotis PI, Durkin D, Storm AL, Hanauer M, Brudno M, Bello SM, Sincan M, Rageth K, Wheeler MT, Oegema R, Lourghi H, Della Rocca MG, Thompson R, Castellanos F, Priest J, Cunningham-Rundles C, Hegde A, Lovering RC, Hajek C, Olry A, Notarangelo L, Similuk M, Zhang XA, Gómez-Andrés D, Lochmüller H, Dollfus H, Rosenzweig S, Marwaha S, Rath A, Sullivan K, Smith C, Milner JD, Leroux D, Boerkoel CF, Klion A, Carter MC, Groza T, Smedley D, Haendel MA, Mungall C, Robinson PN. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res. 2019;47(D1):1018–27. https://doi.org/10.1093/nar/gky1105.
25. Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, Feng G, Kibbe WA. Disease Ontology: A backbone for disease semantic integration. Nucleic Acids Res. 2012;40(D1):940–6. https://doi.org/10.1093/nar/gkr972.
26. Hicks A, Miller MA, Stoeckert C, Mowery D. The Hypertension Ontology [Internet]. Zenodo. 2019. [cited 2021 Mar 30]. https://zenodo.org/record/2605329.
27. Hoehndorf R, Slater L, Schofield PN, Gkoutos GV. Aber-OWL: A framework for ontology-based data access in biology. BMC Bioinformatics. 2015;16(1):26. https://doi.org/10.1186/s12859-015-0456-9.
28. Kalfoglou Y, Schorlemmer M. Ontology mapping: the state of the art. The knowledge engineering review. Vol 18. Cambridge University Press; 2003. p. 131.
29. Hill DP, Adams N, Bada M, Batchelor C, Berardini TZ, Dietze H, Drabkin HJ, Ennis M, Foulger RE, Harris MA, Hastings J, Kale NS, de Matos P, Mungall CJ, Owen G, Roncaglia P, Steinbeck C, Turner S, Lomax J. Dovetailing biology and chemistry: integrating the Gene Ontology with the ChEBI chemical ontology. BMC Genomics. 2013;14(1):513. https://doi.org/10.1186/1471-2164-14-513.
30. Shefchek KA, Harris NL, Gargano M, Matentzoglu N, Unni D, Brush M, Keith D, Conlin T, Vasilevsky N, Zhang XA, Balhoff JP, Babb L, Bello SM, Blau H, Bradford Y, Carbon S, Carmody L, Chan LE, Cipriani V, Cuzick A, Rocca MD, Dunn N, Essaid S, Fey P, Grove C, Gourdine J-P, Hamosh A, Harris M, Helbig I, Hoatlin M, Joachimiak M, Jupp S, Lett KB, Lewis SE,

McNamara C, Pendlington ZM, Pilgrim C, Putman T, Ravanmehr V, Reese J, Riggs E, Robb S, Roncaglia P, Seager J, Segerdell E, Similuk M, Storm AL, Thaxon C, Thessen A, Jacobsen JOB, McMurry JA, Groza T, Köhler S, Smedley D, Robinson PN, Mungall CJ, Haendel MA, Munoz-Torres MC, Osumi-Sutherland D. The Monarch Initiative in 2019: An integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res. 2020;48(D1):704–15. https://doi.org/10.1093/nar/gkz997.

31. Ganzinger M, He S, Breuhahn K, Knaup P. On the ontology based representation of cell lines. PloS ONE. 2012;7(11):48584. https://doi.org/10.1371/journal.pone.0048584.

32. Hamilton CM, Strader LC, Pratt JG, Maiese D, Hendershot T, Kwok RK, Hammond JA, Huggins W, Jackman D, Pan H, Nettles DS, Beaty TH, Farrer LA, Kraft P, Marazita ML, Ordovas JM, Pato CN, Spitz MR, Wagener D, Williams M, Junkins HA, Harlan WR, Ramos EM, Haines J. The PhenX Toolkit: get the most from your measures. Am J Epidemiol. 2011;174(3):253–60. https://doi.org/10.1093/aje/kwr193.

33. Lin FP-Y, Groza T, Kocbek S, Antezana E, Epstein RJ. The Cancer Care Treatment Outcomes Ontology (CCTO): A computable ontology for profiling treatment outcomes of patients with solid tumors. J Clin Oncol. 2017;35(15_suppl):18137. https://doi.org/10.1200/JCO.2017.35.15_suppl.e18137.

34. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu W-L, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform. 2007;40(1):30–43. https://doi.org/10.1016/j.jbi.2006.02.013.

35. Hoehndorf R, Schofield PN, Gkoutos GV. PhenomeNET: a whole-phenome approach to disease gene discovery. Nucleic Acids Res. 2011;39:e119e119.

36. Johnson AEW, Pollard TJ, Shen L, Lehman L.-w. H., Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3(1):1–9. https://doi.org/10.1038/sdata.2016.35.

37. The Apache Groovy programming language [Internet]. [cited 2020 Jan 27]. http://groovy-lang.org.

38. Manning CD, Surdeanu M, Bauer J, Finkel JR, Bethard S, McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit; 2014. p. 5560. [cited 2020 Jan 27]. https://www.aclweb.org/anthology/P14-5010.

39. Gkoutos GV, Schofield PN, Hoehndorf R. The neurobehavior ontology: An ontology for annotation and integration of behavior and behavioral phenotypes. In: International Review of Neurobiology. Elsevier; 2012. p. 69–87.

40. RDF Schema 1.1 [Internet]. [cited 2020 Jan 15]. https://www.w3.org/TR/rdfschema/#ch_label.

41. SKOS Core Vocabulary Specification [Internet]. [cited 2020 Feb 3]. https://www.w3.org/TR/swbp-skos-core-spec/#altLabel.

42. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene Ontology: Tool for the unification of biology. Nat Genet. 2000;25(1):25–9. https://doi.org/10.1038/75556.

43. Slater LT, Bradlow W, Motti DFA, Hoehndorf R, Ball S, Gkoutos GV. A fast, accurate, and generalisable heuristic-based negation detection algorithm for clinical text. Comput Biol Med. 2021;130:104216.

44. Harispe S, Ranwez S, Janaqi S, Montmain J. The semantic measures library and toolkit: Fast computation of semantic similarity and relatedness using biomedical ontologies. Bioinformatics. 2014;30(5): 740–2. https://doi.org/10.1093/bioinformatics/btt581.

45. Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comput Graph Stat. 1996;5:299314.

Publisher's Note