Exploring Binary Relations for Ontology Extension and Improved Adaptation to Clinical Text

Luke T Slater[1,2,6*], Robert Hoehndorf[7], Andreas Karwath[1,2,6], Georgios V Gkoutos[1,2,3,4,5,6]

1 Institute of Cancer and Genomic Sciences, College of Medical and Dental Sciences, University of Birmingham
2 Institute of Translational Medicine, University Hospitals Birmingham, NHS Foundation Trust
3 NIHR Experimental Cancer Medicine Centre
4 NIHR Surgical Reconstruction and Microbiology Research Centre
5 NIHR Biomedical Research Centre
6 MRC Health Data Research UK (HDR UK) Midlands
7 Computer, Electrical and Mathematical Sciences & Engineering Division, Computational Bioscience Research Center, King Abdullah University of Science and Technology

* l.slater.1@bham.ac.uk

Abstract

Background The controlled domain vocabularies provided by ontologies make them an indispensable tool for text mining. Ontologies also include semantic features in the form of taxonomy and axioms, which make annotated entities in text corpora useful for semantic analysis. Extending those semantic features may improve performance for characterisation and analytic tasks. Ontology learning techniques have previously been explored for novel ontology construction from text, though most recent approaches have focused on literature, with applications in information retrieval or human interaction tasks. We hypothesise that extension of existing ontologies using information mined from clinical narrative text may help to adapt those ontologies such that they better characterise those texts, and lead to improved classification performance.

Results We develop and present a framework for identifying new classes in text corpora, which can be integrated into existing ontology hierarchies. To do this, we employ the Stanford Open Information Extraction algorithm and integrate its implementation into the Komenti semantic text mining framework. To identify whether our approach leads to better characterisation of text, we present a case study, using the method to learn an adaptation to the Disease Ontology using text associated with a sample of 1,000 patient visits from the MIMIC-III critical care database. We use the adapted ontology to annotate and classify shared first diagnosis on patient visits with semantic similarity, revealing an improved performance over use of the base Disease Ontology on the set of visits the ontology was constructed from. Moreover, we show that the adapted ontology also improved performance for the same task over two additional unseen samples of 1,000 and 2,500 patient visits.

Conclusions We report a promising new method for ontology learning and extension from text. We demonstrate that we can successfully use the method to adapt an existing ontology to a textual dataset, improving its ability to characterise the dataset, and leading to improved analytic performance, even on unseen portions of the dataset.

Background

Biomedical ontologies provide sets of natural language labels and synonyms that constitute controlled domain vocabularies. These vocabularies are a fundamental resource for text mining [1]. However, ontologies also contain a wealth of semantic features, which enables their use in ontology-based analysis [2]. Recent approaches are increasingly bridging the gap between text mining, semantic analysis, and machine learning [3]. While many recent efforts have focused on the best ways to leverage the axioms that exist in ontologies to improve the power of analyses that use them, less attention has been given to whether text mining could be used to create additional classes and relationships in existing ontologies, particularly extending them in ways that could benefit analysis.

Ontology learning from text, however, has been an area of interest for many years [4,5]. Most approaches focus on learning from literature, and their most common methods of validation are manual, or through use of consistency and satisfiability checks performed by automated reasoners [6]. Other evaluations include application integration, information retrieval, human interaction and performance against gold standard question-answer database [5]. However, the role of ontology learning for adaptation and extension of existing ontologies for improvement of classification tasks has not been previously explored, to our knowledge.

Previous approaches to ontology learning have used Stanford CoreNLP [7] for pre-processing and part-of-speech tagging [8,9]. However, CoreNLP also includes an Open Information Extraction (OpenIE) algorithm, which can extract binary relations from text [10]. The relations it produces are derived from the syntax of sentences, often using the verb of the sentence as a predicate between two nouns or noun phrases which form a subject and object [11]. Some previous investigations have built OWL ontologies from literature text using these relations in combination with topic modelling approaches [12,13]. However, these approaches were only applied to literature text, and have only been used to support applications of information retrieval and literature search. Furthermore, they focus on producing novel ontologies for a particular topic, rather than extending existing ontologies.

One study of using openIE for knowledge graph construction identified several open problems in the domain [14]:

Entity linking Linking named entities in the binary relation to identified entities in a knowledge graph.

Entity Selection Selecting the most optimal and relevant entity amongst candidates produced. For example, which substring of "aorta descent consistent with dissection" to use as the entity in the relation.

Property linking Similar to the problem of entity linking, linking the predicate in the binary relation to an ontology resource.

Representation How to model more expressive binary relations, extrapolating complex relations to several triples.

Instead of using the binary relationships extracted from text to construct an ontology, our approach uses an analysis of the components of extracted binary relationships to identify new classes that can be linked to existing ontology classes. This renders our approach relevant and applicable to the entity linking and entity selection problems. By adding new terms to an existing ontology in this fashion, we hypothesise that the ontology can be adapted to a text corpus, leading to its better characterisation through the annotation of more specific concepts. To explore this hypothesis, we developed a novel approach to ontology extension using triples produced by the OpenIE algorithm. We take a fully automated approach, with no manual curation or evaluation of correctness before final use, with the aim of producing an application ontology that is adapted to a particular dataset, and whose success can be measured by its role in improved classification performance. We apply the approach to a text corpus associated with MIMIC-III patient visits and generate an extended version of the Disease Ontology (DO) [15]. We describe the ontology, and then investigate its use for improving the performance of diagnosis extraction and classification from clinical text.

The triple mining and ontology construction modules are available as part of the Komenti semantic text mining framework, which is available under an open source licence at `https://github.com/reality/komenti` [16]. Files used for the construction of the MIMIC Adapted Disease Ontology (MADO), and for the automated validation are available at `https://github.com/reality/mado_tripulate`.

Results

Approach

We developed a method of extending existing ontologies with new classes mined from text. The resulting ontology facilitates the identification of more specific concepts expressed in the corpus text, in the context of the existing hierarchy. In order to extend an existing ontology, important entities in the text that are related to, but may extend, classes in the ontology of interest need to be identified. Building a vocabulary of relevant terms from the ontology and using it to identify sentences which contain mentions of ontology classes produces candidate sentences that may concern useful new terms.

This leaves the problem that we need to identify modified ontology terms that we can use to produce new classes. Binary relationship extraction algorithms identify relationships between entities in text. In this case, we consider the Open Information Extraction (OpenIE) algorithm, which uses syntactic analysis to identify entities and relationships. Because OpenIE works on this syntactic basis, and is therefore naive to the actual terms in the ontology, it is able to capture noun phrase entities which contain ontology terms, but extend them (make them more specific), allowing us to identify new classes that extend existing ontology classes. Furthermore, by examining only binary relations that contain at least one ontology term in both its subject and object, we can assume that these entities are both important to the text and relevant to the context of the ontology. The subject and object entity strings can then be decomposed into a combination of ontology terms and an optional specifying component. These new terms can then be added to the ontology, making them descendants of the classes they extend.

Once this has been done, the extended ontology can be used to annotate those more specific concepts in the text corpus. In this investigation, we do not make use of the binary relations themselves to add additional relationships to the ontology, beyond identifying classes to add to the hierarchy. The employment of binary relationship extraction in this case can be considered as a 'filter,' identifying important entities in the text to add to the ontology. Figure 1 describes the overall approach.
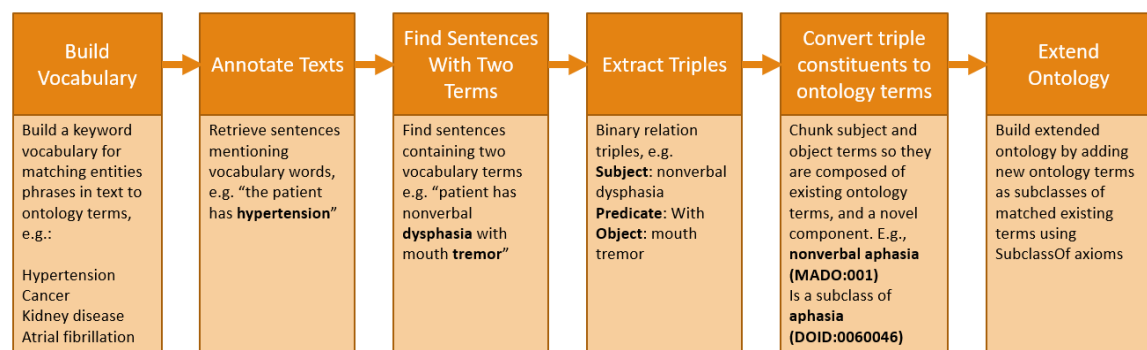

Figure 1. Description of our workflow to extract new classes from a corpus by annotating a text, and examining binary relations.

Vocabulary Construction and Annotation 82

For the vocabulary construction and annotation stages, our approach uses Komenti's existing 83
pipeline for vocabulary construction and text annotation [16]. The vocabulary construction process 84
uses the AberOWL reasoner-based ontology access framework [17] to build a list of labels and 85
synonyms for a given input ontology. The annotation process uses this vocabulary to match 86
mentions of the relevant concepts in text, using the Stanford RegexNER annotator implemented by 87
CoreNLP. This outputs a list of annotated terms and the sentences that contain them. To identify 88
sentences that could contain valid binary relations that pertain to terms in the input ontology, our 89
algorithm then examines the list of annotations, filtering it to contain only sentences which contain 90
at least two ontology terms. 91

Entity Selection and Linking 92

The remaining sentences are then used as input to the Open Information Extraction (OpenIE) 93
algorithm, which extracts binary relationships from input text. OpenIE produces triples, which 94
represent a relationship between two entities in the form of a subject, predicate, and object. Since 95
we are passing it sentences which contain two terms from our ontology, we intend to capture 96
relationships which concern subjects and objects related to our ontology: either equivalent to a term, 97
or containing terms that can be transformed into a new specifying class. 98

To integrate these new entities into our ontology, we need to describe them in terms of existing 99
ontology classes. Subjects and objects are again annotated using the vocabulary that was previously 100
used to identify candidate sentences. Subsequently a representation of the entities are produced by 101
repeatedly identifying the longest annotation corresponding to an ontology term. When these have 102
been exhausted, either the whole string has been consumed, or we are left with an arbitrary 103
specifying string. In this way, we produce a new class, which we assert is a subclass of the existing 104
ontology terms contained in the entity string. The algorithm performing this process is described in 105
Figure 2. 106

For example, an entity label appearing as a subject or object in a triple could be "extreme 107
hypertension." Using a vocabulary constructed from the Disease Ontology (DO), the first annotation 108
employed by the algorithm would match *hypertension* (`DOID:10763`). The algorithm would then 109
recurse, attempting to annotate the remaining unconsumed string, 'extreme.' Since this word does 110
not feature in DO, no annotation can be made. Thus, 'extreme' is noted as an unmatched concept, 111
and we can produce a new class with the label 'extreme hypertension,' as a subclass of *hypertension* 112
(`DOID:10763`). New terms can also be composed of multiple ontology classes, such as for the example 113
'activation of apoptosis cell.' Using a vocabulary constructed from the Gene Ontology, the algorithm 114
would first identify *activation of apoptosis* (`GO:0006915`) as the annotation with the maximal string 115
length. On the second round, it would match *cell* (`GO:0005623`), leading to a new 'activation of 116
apoptosis cell' class, which would be a subclass of both *activation of apoptosis* and *cell*. This can 117
lead to rich hierarchies, as other processed entities may specify further upon these intermediate 118
classes, such as for 'activation of apoptosis cell of mz-cha-1' or 'activation of apoptosis less than 119
control.' If no ontology term is matched at all, the whole string is returned as an unmatched concept, 120
and is later thrown away by the algorithm, as it cannot be integrated with the ontology hierarchy. 121

Ontology Extension 122

For each triple, after the subject and object have been processed, we add both to the list of new 123
classes to add to the ontology, if we were able to construct an ontology term from both the subject 124
and object. Otherwise, it is discarded. This is to ensure that we extend existing ontology structures 125
and features; we would not be able to construct an ontology class from an entity label that does not 126
contain at least one ontology term. Once all triples have been evaluated, a Komenti module is used 127
to create a new ontology expressed by subclass axioms and labels, asserting the new classes in the 128
context of the target ontology. Once this has been done, an import to the original ontology is added. 129

Data: label = Given string label
Result: A new ontology class *term*
def term
while *strlen(label) > 0* **do**
> a = annotate(label)
> **if** *a* **then**
> > a = maxlen(label)
> > newTerm = ontologyClassFromAnnotation(a)
> > label = strreplace(label, a)
>
> **end**
> **else**
> > newTerm = unmatchedOntologyClass()
> > label = ″
>
> **end**
> **if** *term* **then**
> > newTerm.parentTerm = term
>
> **end**
> term = newTerm

end

Figure 2. Algorithm for building hierarchical ontology terms from matched openIE labels by recursively consuming the label. The annotate function returns a list of ontology terms that were found in the input string.

MIMIC Adapted Disease Ontology   130

We applied our method to texts associated with 1,000 patient visits sampled from the MIMIC-III   131
critical care database. Our input vocabulary contained terms from the Disease Ontology (DO) and   132
the Phenotypic Trait Ontology (PATO) [18]. The DO terms were restricted to classes that contained 133
cross references to ICD-9, which is the medical terminology used to annotate MIMIC-III patient   134
visits in the structured data. We did this to reduce the running time of the experiment, by   135
restricting our analysis to classes likely to be represented in the structured outcome data. We will 136
refer to this as the 'Trim DO' vocabulary, to distinguish it from the 'Full DO' vocabulary, which 137
contains all non-obsolete classes in the ontology. This produced the MIMIC Adapted Disease   138
Ontology (MADO). The unmodified DO contains 18.591 classes, while MADO contains 20,818,   139
meaning that the approach produced 2,227 new classes.   140

Diagnosis Classification   141

To evaluate whether MADO could lead to an improved characterisation of the dataset, we compared   142
its performance at classifying shared primary diagnosis over sets of MIMIC patient visits. We used 143
the annotation terms as input to the Resnik semantic similarity algorithm, using the Best Match   144
Average, and using a measure of information content derived from the probability of the annotation 145
appearing in the entire set of annotations. We then evaluated our approach's performance at   146
identifying a shared primary diagnosis when ranking patient visits by their similarity scores.   147

Our evaluation included several settings, spanning different combinations of vocabularies used as 148
input, and ontologies used to calculate the similarity scores. Since a Trim DO vocabulary (the subset 149
with ICD-9 cross mappings) was used to produce MADO, we also included results for use of the   150
Trim DO vocabulary without extensions to ensure that improved performance did not derive from   151
use of the slimmer vocabulary alone.   152

Table 1. Performance for matching first diagnosis of MIMIC patients with different combinations of annotation vocabularies and ontologies used for semantic similarity score calculation.

Setting	Vocabulary	Ontology	MRR	MAP	Precision	p-value ($<$)
Source-1000						
	Full DO	DO	0.939	0.844	0.937	-
	Trim DO	DO	0.947	0.88	0.917	0.01466
	MADO	MADO	**0.995**	**0.916**	**0.992**	6.458e-05
Unseen-1000						
	Full DO	DO	0.931	0.841	0.929	-
	Full DO	MADO	0.88	0.795	0.876	0.007279
	Trim DO	DO	0.958	0.885	0.932	0.0001161
	Trim DO	MADO	0.958	0.887	0.933	0.003315
	MADO	MADO	**0.986**	**0.92**	**0.981**	0.009601
Unseen-2500						
	Full DO	DO	0.943	0.842	0.941	-
	Full DO	MADO	0.896	0.798	0.891	2.2e-16
	Trim DO	DO	0.932	0.865	0.898	0.004658
	Trim DO	MADO	0.933	0.866	0.899	8.035e-05
	MADO	MADO	**0.986**	**0.913**	**0.98**	2.2e-16

Discussion

Our results show that the developed method was successful in constructing a MIMIC adapted disease ontology. The results of the characterisation evaluation demonstrate that the use of the MADO ontology and vocabulary consistently out-performs two different configurations of DO vocabulary used with DO and MADO for classificaiton of primary diagnosis. Furthermore, not only does our method improve characterisation based on text the ontology that was learned upon, it also improves characterisation from unseen text in the same corpus. This provides evidence that this process of training and extending an ontology on a dataset adapts that ontology to it, increasing its ability to characterise the dataset.

One limitation of the study is that it only considers primary diagnosis. Our evaluation could be considered a simplified version of those presented in several other investigations which use various learning methods to assign ICD-9 codes based on MIMIC-III text [19]. Our work only assigns the primary diagnosis, as a method of evaluating improvement between two settings. To make the work more practical, and to further evaluate the approach, we could consider transforming the task to classify several codes. However, this presents several challenges, primarily due to semantic similarity being an unsupervised and unguided approach. However, we believe that the similarity approach should capture this information, e.g. that patients with a secondary diagnosis of 'pain' would be more similar to other patients with a pain diagnosis. The challenge is in developing a decision procedure for making that classification, while overlooking other reasons (i.e. a different shared diagnosis) that patients may be highly similar. The advantages of developing this kind of decision procedure would be the ability to use this method for a more complete characterisation of the dataset, as well as the ability to directly compare performance with other approaches using the same evaluation methods. A follow-up investigation could measure how well annotations predicted overall patient diagnosis structure. These methods could also be applied to other benchmark issues that have typically been solved with similarity-based approaches, such as gene enrichment or protein-protein similarity. A future validation could also work to determine whether larger training sets could lead to a greater validation performance, and whether this value could be established as a hyper-parameter set based on size or other information concerning a target corpus.

Nevertheless, the evaluation showed that our approach improved the performance of the task of classifying primary diagnosis from text. This has potential applications in the clinical space, including automated diagnosis and classification of patients from text profiles. Previous work

involved the use of semantic similarity over the Human Phenotype Ontology for differential diagnosis [184] of medical conditions [20–22]. However, these phenotypic profiles are usually produced by experts, or [185] by patients themselves. We would like to investigate whether it is possible to construct these profiles [186] from text, particularly aided by methods that can construct and adapt multiple ontologies around [187] particular diseases, datasets, and relevant entities. [188]

While our work shows that the approach can adapt an individual ontology to a text corpus for [189] improved classification, this work could potentially have other uses and implications. As well as [190] extending other single ontologies to better characterise other text corpora with application to other [191] classification problems, the approach also has the potential to capture relationships between different [192] domain ontologies. While this investigation only involves DO, PATO, and the classes they include, [193] the algorithm would also work to connect more ontologies that describe different domains. For [194] example, a vocabulary with terms from DO and GO would be capable of producing new classes that [195] combined genetic components and diseases, such as in the case of 'cancer cell.' This would [196] potentially allow for the integration of information from different domains contributing to singular [197] tasks. In the case of semantic similarity, involving several ontologies produces an additional level of [198] complexity [23]. [199]

Furthermore, while the approach employs the use of binary relation extraction, it does not [200] currently make use of the relationships themselves. Observed relationships from text could be used [201] as ABox axioms, defining observations of instances of ontology classes in text. Semantic similarity [202] approaches do not make use of such information, although recent approaches to machine learning [203] with ontologies, such as OPA2Vec [24] could take these relationships into account. This would [204] potentially further increase the power of cross-ontology and integrated domain analysis, as [205] relationships could be established between classes of different ontologies through observations as well [206] as class composition. [207]

Both the construction of the ontology, and thereby analysis performance could be affected by [208] investigating the use of other Komenti features, particularly those for negation detection and [209] synonym expansion [25, 26]. In the former case, the evaluation of negation could prevent incorrect or [210] explicitly negated facts from being used in the produced knowledgebase, while the latter case has [211] been shown to improve overall characterisation of text. [212]

Conclusions [213]

Our work describes a new approach to extending existing ontologies with new classes through [214] analysis of text corpora. The ontology extensions essentially adapt the ontology to better suit the [215] text corpus, and leads to improved classification performance using the extended set of annotations, [216] even on unseen portions of the dataset. We believe that the approach is promising, and could have [217] further implications, which we intend to explore in future work. [218]

Methods [219]

The algorithm was developed using the Komenti semantic text mining framework, which is available [220] under an open source licence at `http://github.com/reality/Komenti` [16]. It makes use of [221] Stanford CoreNLP [7], and particularly the RegexNER annotator for identifying entities in text, and [222] the openIE algorithm for extracting triples from text. [223]

Corpus Generation [224]

Texts were sourced from the MIMIC-III critical care database [27]. MIMIC is a publicly available [225] dataset describing over 50,000 critical care visits, including natural language text linked with many [226] structured features. We sampled three sets of patient visits from MIMIC: one of 1,000 patient visits [227] for learning the ontology, and test sets of 1,000 and 2,500 patient visits. Patient visits were [228]

associated with their primary diagnosis, the first listed for the record in the ICD9_DIAGNOSES   229
table. These codes are produced by clinical coding specialists, by examining the texts associated   230
with the visit. We then concatenated all text records for each patient visit from the NOTEEVENTS   231
table into one text file, and pre-processed the text to remove newlines, improve sentence delineation,   232
and lemmatise words.   233

   We used two base vocabularies for our experiments. The first, referred to as 'Full DO'   234
throughout the article, contained all labels and synonyms in the Disease Ontology (DO) [15]. The   235
second was limited to labels and synonyms in 2,118 DO terms that were annotated with a database   236
cross-reference to ICD-9 used in an effort to reduce noise from terms not represented in ICD-9. We   237
obtained the unexpanded and expanded synonyms for these terms on 08/07/2020. Both sets of   238
labels were also lemmatised.   239

Ontology Extension   240

Our ontology extension approach was applied to the pre-selected training set of 1,000 patient visits,   241
using the Trim DO vocabulary. The produced ontology was then merged with DO, by copying all   242
axioms from DO into the new ontology. This merged ontology is referred to as the MIMIC Adapted   243
Disease Ontology (MADO). The ontology constructor also contains an optional flag that outputs a   244
new Komenti-compatible vocabulary with the novel classes and pre-existing ontology classes used as   245
components in the ontology extension. This feature was used to create a new vocabulary, which was   246
concatenated with the Trim DO vocabulary, and used in the experiments that refer to the MADO   247
vocabulary.   248

Evaluation   249

Komenti was used to annotate the text associated with each patient visit. We then used the set of   250
terms associated with it to produce a semantic similarity matrix for patient visits, using the Resnik   251
measure of pairwise similarity for each annotated term [28], normalised into a groupwise measure   252
using the best match average method [29]. Information content was calculated using the probability   253
of the term appearing as an annotation in the totality of the set of annotations [28]. The similarity   254
matrix was computed using the Semantic Measures Library [30].   255

   We evaluated the similarity matrix using mean reciprocal rank, mean average precision, and   256
precision to measure performance in predicting shared primary patient diagnosis. A true case was   257
considered to be whether a pair of patient visits had the same primary diagnosis (as per the   258
MIMIC-III database). For mean average precision, we considered only the 10 most similar patients   259
for each patient, establishing a cut-off to limit the effect of patients who were ranked more similar   260
for reasons of other diagnosis. The p-value was calculated using the built-in *wilcox.test* function of R   261
version 3.4.4 [31].   262

Competing interests   263

The authors declare that they have no competing interests.   264

Author's contributions   265

LTS conceived of the study, performed the experiments, implemented the software, and wrote the   266
first draft of the manuscript. LTS, RH, and AK contributed to the experimental design. RH and   267
GVG contributed to the manuscript. GVG and RH supervised the project. All authors contributed   268
to the interpretation of the results, as well as revised and approved the manuscript for submission.   269

Acknowledgements

Funding

References

1. Spasic, I., Ananiadou, S., McNaught, J., Kumar, A.: Text mining and ontologies in biomedicine: Making sense of raw text. Briefings in Bioinformatics **6**(3), 239–251 (2005). doi:10.1093/bib/6.3.239

2. Hoehndorf, R., Schofield, P.N., Gkoutos, G.V.: The role of ontologies in biological and biomedical research: A functional perspective. Briefings in Bioinformatics **16**(6), 1069–1080 (2015). doi:10.1093/bib/bbv011

3. Kulmanov, M., Smaili, F.Z., Gao, X., Hoehndorf, R.: Machine learning with biomedical ontologies. bioRxiv, 2020–0507082164 (2020). Chap. New Results. doi:10.1101/2020.05.07.082164

4. Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications, 2006 edition edn. Springer, New York, N.Y (2006)

5. Asim, M.N., Wasim, M., Khan, M.U.G., Mahmood, W., Abbasi, H.M.: A survey of ontology learning techniques and applications. Database **2018** (2018). doi:10.1093/database/bay101

6. Konys, A.: Knowledge Repository of Ontology Learning Tools from Text. Procedia Computer Science **159**, 1614–1628 (2019). doi:10.1016/j.procs.2019.09.332

7. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60 (2014). doi:10.3115/v1/P14-5010

8. Petit, J., Boisson, J.-C., Rousseaux, F.: Discovering cultural conceptual structures from texts for ontology generation. In: 2017 4th International Conference on Control, Decision and Information Technologies (CoDIT), pp. 0225–0229 (2017). doi:10.1109/CoDIT.2017.8102595

9. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL '03, vol. 1, pp. 423–430. Association for Computational Linguistics, Sapporo, Japan (2003). doi:10.3115/1075096.1075150

10. Angeli, G., Johnson Premkumar, M.J., Manning, C.D.: Leveraging Linguistic Structure For Open Domain Information Extraction. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 344–354. Association for Computational Linguistics, Beijing, China (2015). doi:10.3115/v1/P15-1034

11. Roller, S., Kiela, D., Nickel, M.: Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora. arXiv:1806.03191 [cs] (2018). 1806.03191

12. Nagulapati, R., Chandrashekar, M., Lee, Y.: Transformation from Publications to Diabetes Ontology using Topic-based Assertion Discovery. In: 2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W), pp. 9–18 (2018). doi:10.1109/ICHI-W.2018.00009

13. Nagabhushan, M., Nagulapati, R., Chandrashekar, M., Lee, Y.: Constructing dynamic ontologies from biomedical publications. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 581–584 (2017). doi:10.1109/BIBM.2017.8217715

14. Martinez-Rodriguez, J.L., Lopez-Arevalo, I., Rios-Alvarado, A.B.: OpenIE-based approach for Knowledge Graph construction from text. Expert Systems with Applications **113**, 339–355 (2018). doi:10.1016/j.eswa.2018.07.017

15. Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.-W.W., Mazaitis, M., Felix, V., Feng, G., Kibbe, W.A.: Disease Ontology: A backbone for disease semantic integration. Nucleic Acids Research **40**(D1), 940–946 (2012). doi:10.1093/nar/gkr972

16. Slater, L.T., Bradlow, W., Hoehndorf, R., Motti, D.F., Ball, S., Gkoutos, G.V.: Komenti: A semantic text mining framework. bioRxiv, 2020–0804233049 (2020). Chap. New Results. doi:10.1101/2020.08.04.233049

17. Hoehndorf, R., Slater, L., Schofield, P.N., Gkoutos, G.V.: Aber-OWL: A framework for ontology-based data access in biology. BMC Bioinformatics **16**(1), 26 (2015). doi:10.1186/s12859-015-0456-9

18. Gkoutos, G.V., Mungall, C., Dolken, S., Ashburner, M., Lewis, S., Hancock, J., Schofield, P., Kohler, S., Robinson, P.N.: Entity/quality-based logical definitions for the human skeletal phenome using PATO. In: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 7069–7072 (2009). doi:10.1109/IEMBS.2009.5333362

19. Gehrmann, S., Dernoncourt, F., Li, Y., Carlson, E.T., Wu, J.T., Welt, J., Foote Jr., J., Moseley, E.T., Grant, D.W., Tyler, P.D., Celi, L.A.: Comparing Rule-Based and Deep Learning Models for Patient Phenotyping. arXiv:1703.08705 [cs, stat] (2017). 1703.08705

20. Girdea, M., Dumitriu, S., Fiume, M., Bowdin, S., Boycott, K.M., Chénier, S., Chitayat, D., Faghfoury, H., Meyn, M.S., Ray, P.N., So, J., Stavropoulos, D.J., Brudno, M.: PhenoTips: Patient Phenotyping Software for Clinical and Research Use. Human Mutation **34**(8), 1057–1065 (2013). doi:10.1002/humu.22347

21. Köhler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dölken, S., Ott, C.E., Mundlos, C., Horn, D., Mundlos, S., Robinson, P.N.: Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. The American Journal of Human Genetics **85**(4), 457–464 (2009). doi:10.1016/j.ajhg.2009.09.003

22. Köhler, S., Øien, N.C., Buske, O.J., Groza, T., Jacobsen, J.O.B., McNamara, C., Vasilevsky, N., Carmody, L.C., Gourdine, J.P., Gargano, M., McMurry, J.A., Danis, D., Mungall, C.J., Smedley, D., Haendel, M., Robinson, P.N.: Encoding Clinical Data with the Human Phenotype Ontology for Computational Differential Diagnostics. Current Protocols in Human Genetics **103**(1), 92 (2019). doi:10.1002/cphg.92

23. Solé-Ribalta, A., Sánchez, D., Batet, M., Serratosa, F.: Towards the estimation of feature-based semantic similarity using multiple ontologies. Knowledge-Based Systems **55**, 101–113 (2014). doi:10.1016/j.knosys.2013.10.015

24. Smaili, F.Z., Gao, X., Hoehndorf, R.: OPA2Vec: Combining formal and informal content of biomedical ontologies to improve similarity-based prediction. arXiv:1804.10922 [cs] (2018). 1804.10922

25. Slater, L.T., Bradlow, W., Motti, D.F., Hoehndorf, R., Ball, S., Gkoutos, G.V.: A fast, accurate, and generalisable heuristic-based negation detection algorithm for clinical text. bioRxiv, 2020–0703187054 (2020). Chap. New Results. doi:10.1101/2020.07.03.187054

26. Slater, L.T., Bradlow, W., Ball, S., Hoehndorf, R., Gkoutos, G.V.: Improved characterisation of clinical text through ontology-based vocabulary expansion. bioRxiv, 2020–0710197541 (2020). Chap. New Results. doi:10.1101/2020.07.10.197541

27. Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.-w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific Data **3**(1), 1–9 (2016). doi:10.1038/sdata.2016.35

28. Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. arXiv:cmp-lg/9511007 (1995). cmp-lg/9511007

29. Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., Chen, C.-F.: A new method to measure the semantic similarity of GO terms. Bioinformatics **23**(10), 1274–1281 (2007). doi:10.1093/bioinformatics/btm087

30. 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 **30**(5), 740–742 (2014). doi:10.1093/bioinformatics/btt581

31. Ihaka, R., Gentleman, R.: R: A language for data analysis and graphics. Journal of computational and graphical statistics **5**(3), 299–314 (1996)