Formal Representations of Ontologies for Automation of Analyses

J. Hastings, 6 September 2021

Abstract

Ontologies are a form of symbolic logic-based knowledge representation technology in which relevant content, in the form of entities and their definitions, is represented in computable form. They are widely used throughout the biomedical sciences, and are starting to be used in the behavioral sciences. They provide a structured, computable representation of the entities and interrelationships in a domain of interest that can be used to drive multiple computational applications in data management and scientific research. Ontologies are represented in logical languages and are part of the family of semantic technologies – technologies for formally representing meanings. Recently, ontologies are starting to be used together with variants of modern machine-learning-based artificial intelligence in order to drive increasingly sophisticated applications that are able to harness automation of knowledge, learning and inference together. This paper provides an overview of the formal representation of ontologies in ways that support automated reasoning with such representations, and the application of such ontologies together with large-scale sources of data.

Table of Contents

Introduction

All scientific domains accumulate a wealth of knowledge about the entities that are the subject of investigations and observations within that domain, and the way that such entities are interrelated. Knowledge must be structured in order to be exchanged and interpreted. As technology advances, larger and larger collections of more and more different types of data are being amassed, stored, analyzed, interpreted, exchanged and integrated. The proper management of such datasets requires that the meaning of the data be clear. At the same time, volumes of scientific research outputs – primarily in the form of written natural language reports including embedded data tables – are increasing exponentially as the scientific research enterprise accelerates on all fronts. Organizing and integrating the findings from these reports also requires computational support, ideally structured around the very same entities as are mentioned in the underlying data on which reports are based, thereby strengthening the chain of accountability and transparency through all discovery stages, from raw data through analysis and output to reporting.

The development and adoption of standards for the entities that form the subject matter of a given domain has a long history across scientific disciplines tracing back to early categorization efforts by ancient philosophers, with the periodic table in chemistry a prominent example from more recent history that is still used and developed today. *Ontologies,* while being rooted in historical developments, have developed as a particularly modern incarnation of a classification system, arising from the *digitalization* of scientific research and the concomitant need for *data management* (Hastings, 2017) and the analysis of large-scale data resources.

What are ontologies?

There are many answers to the question of what ontologies are (Neuhaus, 2018), but in their essence ontologies are *representational documents* that contain formal, standardized descriptions of types of thing. They build on and extend other types of standardized representation of types of thing including controlled vocabularies and terminologies, hierarchies and taxonomies, and supplement them with semantic relationships and logical definitions that support consistency checking, error detection and automated inferences (Figure 1).

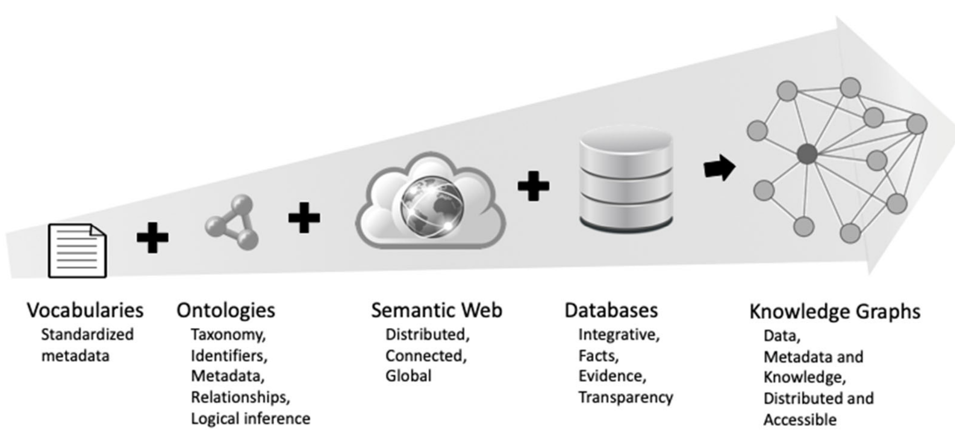

Figure 1: Ontologies in the wider context of knowledge representation technologies

Ontologies usually cover largely definitional knowledge, that is, *invariant* aspects of entities (Rector *et al.*, 2019), as far as possible. Ontologies are one form of computable representation structure among others, including linguistic resources, databases which store evidence and facts, and knowledge resources that contain types of knowledge other than descriptions, such as knowledge about *associations, rules* or *causal connections*. One of the benefits of ontologies is that they provide an index of unambiguous identifiers for the entities that they define which can be used to uniquely and unambiguously refer to the same types of entity in other formalisms as well, even those which capture other sorts of knowledge, and thereby an ontology serves as a hub allowing different sorts of knowledge to be integrated *semantically*, that is, integrated by virtue of being *about the same thing* – a powerful capability in the typical modern data-rich research context.

What do ontologies contain?

Ontologies consist of several basic elements. The fundamental unit of structure in an ontology is an *entity*, corresponding to the sorts of things that are the subjects of scientific investigations. These usually reflect general types of thing – i.e. capturing repeatable features of the world such as are usually the subject of scientific investigation – rather than specific, unique, individual things. Examples of entities include *human being*, *smoking*, *emotion, income, policy* and so on. Entities corresponding to general types of thing may be referred to as classes, while entities corresponding to specific individuals may be referred to as instances.

Organized around the structural units of the entities, ontologies also contain a wealth of metadata (Figure 2). The representation of an entity in an ontology is associated with a name, a definition that should specify as far as possible unambiguously the nature of the entity being represented (Michie, West and Hastings, 2019), an identifier, and other relevant metadata including synonyms and examples of usage. Identifiers are unique for each entity and are usually scoped within a particular *namespace*, to which the ontology belongs. For example, the Behaviour Change Intervention Ontology has namespace BCIO and identifiers which take the form BCIO:xxxxxx where xxxxxxx is a unique number.

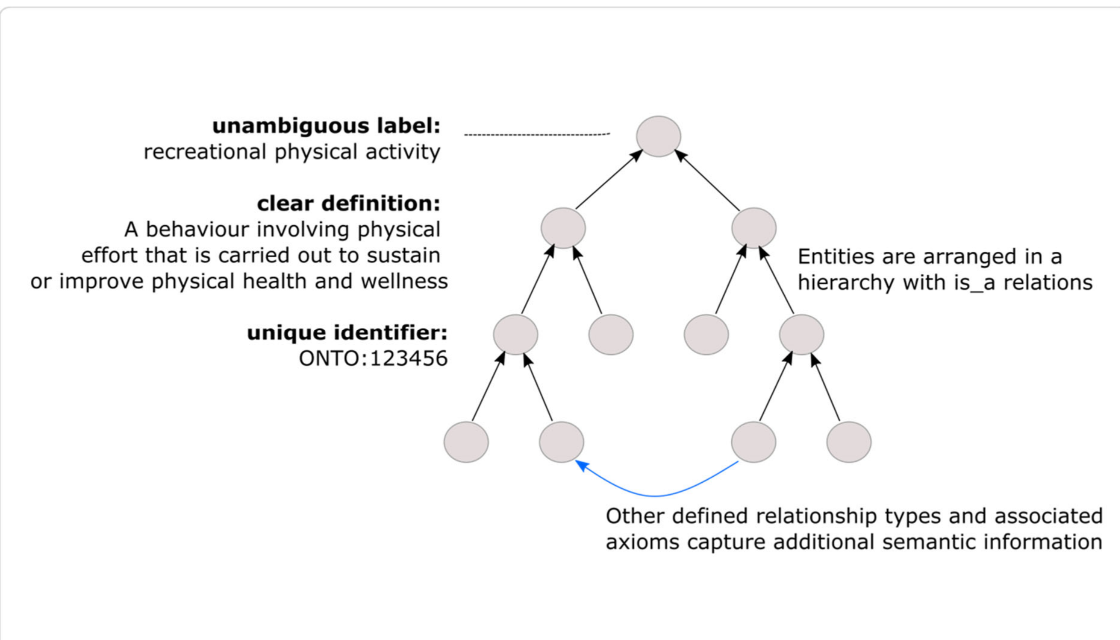

Figure 2: A schematic illustration of an ontology class hierarchy indicating class-level associated metadata.

It is considered best practice for identifiers to be semantics-free (i.e. not to contain the class name or definition) for two reasons: 1) to support annotations in multiple language contexts so as to provide global reach and accessibility, and 2) in order to promote stability even as scientific knowledge and the accompanying ontology representation evolve. Ontology providers must commit to maintaining identifiers for the long term, so that if they are used in annotations or other application contexts the user can still rely on them. As an ontology evolves, it is possible for multiple entries to be merged into one, but in these cases alternate identifiers will still be maintained as secondary identifiers. When a class is deemed to no longer be needed within the ontology it may be marked as obsolete, which then indicates that the ID should not be used in further annotations, although it is preserved for historical reasons.

In addition to metadata, the content in an ontology is interrelated in a hierarchy or *taxonomy* from the general to the specific, which allows different sorts of things to be delineated, described and annotated at multiple different levels of aggregation or grouping. For example, an ontology hierarchy might capture the following chain of specializations of activity descriptions: *behaviour—physical activity—recreational physical activity—yoga activity.* Hierarchy provides organization which is useful for browsing and querying data, and moreover allows aggregation and clustering at multiple different levels of generality, allowing data-driven analyses to test for meaning and patterns in large datasets at multiple different grouping levels.

In addition to the hierarchical classification, which is represented in the formal structure of the ontology by *is a* (also known as *subClassOf*) relations, further types of relationships can be semantically specified to capture all the different ways that entities can be related. Examples of semantic relationships commonly used in ontologies across domains include parthood (*is part of* and *has part*), participation (*participant of*), and dependence (*inheres in*).

Types of ontologies and their applications

Ontologies in the modern sense originally arose from research in computer science. In particular, in the early days of research into artificial intelligence, it was recognized that for artificial systems to act intelligently, they would need to have a representation of aspects of the common-sense structure of the world so as to reason about and interact with their physical environments. This encompassed such entities and domains as objects, temporality and causality (McCarthy and Hayes, 1969; Hayes, 1989), and gave rise to the field of 'knowledge representation' as a sub-discipline of computer science. Knowledge representation is still a burgeoning field of research in computational science, and from it have emerged many different languages and techniques for encoding structured knowledge in computable form, and algorithms for deriving inferences from such knowledge. While the objective of representing in computable form everything we know about the common-sense structure of the world and reasoning with that proved to be a task much more challenging than its early proponents anticipated (Davis and Marcus, 2015), nevertheless from this field modern ontologies have inherited many tools and technologies, including the widely used Web Ontology Language (OWL) (Horrocks *et al.*, 2007) that was designed to support large-scale deployment of ontologies in the Semantic Web.

In a somewhat parallel development, at the turn of the century, the sequencing of the human genome and the accompanying technological revolution in the biological sciences gave rise to massive quantities of new types of data that needed organization, integration, and to be assigned meaning. For example, in order to be able to consistently integrate data about the biological processes that genes were involved in across data arising from a range of different types of study on a range of different types of animal, a common vocabulary for molecular functions and biological processes was created – the *Gene Ontology,* or GO (Ashburner *et al.*, 2000). A wide range of additional ontologies were developed in quick succession within the biological and biomedical sciences, to address exploding quantities of data. Important to these ontologies, commonly known as "bio-ontologies", was that they contained a comprehensive catalogue of the sorts of entities in the domain, down to a quite detailed level of description. For example, the Gene Ontology currently includes 43,878 classes. On the other hand, their logical and semantic structure was often simple – just using a small number of semantic relationships – and the logical language that was used for their formalization was for the most part not very expressive.

The ontologies of today have developed from these diverse influences into a thoroughly integrative technology, and as such they now bring together elements arising from across their wide range of different disciplines of origin. They can nevertheless be grouped into families or traditions of practice depending on their approach to capturing content and the different emphasis placed on their different aspects. Some ontologies are more focused on specifying the broad regularities and logical structure of entities in a given domain, with a focus on complex logical and structural formalizations for a few entities rather than catalogue-like classifications of many entities. These more 'structurally' oriented ontologies can be seen as having the objective of driving software applications that can harness, and need to be aware of, that structure. For example, the emotion ontology developed in (Berthelon and Sander, 2013) to support a use case in affective computing contains only 16 classes, with the main purpose of the ontology being to represent the broad structure of emotions through the relationships between emotion and valence, emotion and arousal, and the context of emotion. While a selection of emotions are included in the ontology (e.g., anger, sadness, fear), the purpose is clearly not to comprehensively list types of emotion. On the other hand, an alternative emotion ontology (Hastings *et al.*, 2011) that was developed to annotate scientific studies of emotions and related phenomena, contains over 600 classes, amongst which close to 60 are different types of emotion. Nevertheless, while we may draw a distinction between these two extremes, most modern ontologies fall somewhere in between.

Within ontologies developed with a scientific purpose, that is, a purpose tied to objectives within empirical research in a given domain, as opposed to ontologies that are developed with technological purposes within a particular engineering context, there are some that are aiming to serve as a single point of reference for knowledge about an aspect of the whole domain – they are *reference* ontologies. Historically, the Gene Ontology was one of the first examples of a reference ontology in the biomedical domain, but subsequently many others have been developed. The OBO Foundry (Smith *et al.*, 2007) is an organization that arose out of the need to coordinate and interconnect different reference ontologies covering different aspects of the biomedical domain in such a way that they could be used together in pipelines and analyses. It provides (a) a portal where such ontologies can be exchanged, (b) a set of best practices and guidelines for their development, both procedurally and technically, and (c) a suite of common infrastructural utilities for the development of ontologies following those guidelines. Important to the principles for Foundry ontologies is that they don't overlap – that each ontology covers its own unique portion of an overall domain in such a way that overlapping content is minimized, which necessitates that when ontologies do need to cover the same conceptual ground, they coordinate and find a way to re-use or import the content from one to another. The OBO Library collection[1] – consisting of ontologies that aim to follow the Foundry guidelines – currently consists of more than 100 ontologies. We can furthermore distinguish between ontologies developed for a particular domain (domain ontologies) and those which are developed to cross domains, for example, the Basic Formal Ontology (Arp,

[1] Listed at http://www.obofoundry.org/.

Smith and Spear, 2015) that serves as the highest-level classification for many of the Foundry ontologies, is specifically designed to provide definitions for semantic types that cut across individual domains, such as the distinction between objects and processes.

In contrast to reference ontologies, ontologies that are developed with more of a focus on a specific application or purpose are often called *application* ontologies. Such application ontologies are driven not by a vision of comprehensively mapping out the content of a certain domain but rather by the need for supporting a specific use case or software application. In scientific contexts alone there are several use cases for ontologies that arise from the pervasive needs for data management and analysis. For example, ontologies can structure the content in databases and ensure common frameworks for data in support of data exchange. They can be used to annotate data and text in such a way that allows those data and text to be unambiguously interpreted subsequently. And they support the development of neat user interfaces through their hierarchies and axioms, which user interfaces may offer browsing, searching or querying that is enhanced by the ontology. A particularly interesting application of ontologies to large-scale (“big”) annotated datasets is a form of statistical analysis known as over- or under-representation analysis. For this, the assumption that the ontology covers the full scope of the range of values that can be expected in the data is key. The ontology is then used to explore the distribution of the annotated values to ontology classes – answering the question ‘does the distribution of annotations match what would be expected by chance, given some additional knowledge about the background distribution of values in this type of data?’ It is this capability that has been one of the main drivers for the pervasive adoption of the GO and similar bio-ontologies in data analysis pipelines, because it allows for making sense of huge datasets in terms by determining the ontology categories that are more, or less, represented in them (Bauer, 2017), as for example we may determine that molecular repair operations are over-represented in tissues sampled from the brains of people with neurodegenerative disorders as compared to those of healthy controls.

Nevertheless, to some extent, most modern ontologies strive to support multiple applications at the same time, and the distinction between reference and application ontologies is therefore not always clear. This is because it is hard and time consuming to develop ontologies and thus it makes sense if invested resources are capitalized on to the greatest extent possible. Indeed, far too many ontologies are developed and then not re-used for one reason or another. Reusability – as measured in practice through adoption – is one of the criteria that can be considered a metric for the success of an ontology. Metrics for the success or evaluation of ontologies is the subject of the next section.

Evaluating ontologies

Informally, there are a number of criteria that can be used to evaluate how good, fit for purpose, or correct an ontology is (Brank, Grobelnik and Mladenic, 2005; Obrst *et al.*, 2007). These reflect various different dimensions of quality. The first dimension of quality is intrinsic. For example, we might evaluate if the ontology is clear, concise, and well-organized. The second dimension of quality relates to the technical implementation. We might ask whether the ontology is logically sound (i.e. contains no contradictions), technically correct, and even if it is computationally elegant, that is, if it expresses itself succinctly and making good use of the underlying logical language. The third dimension of quality relates to the relationship between the ontology itself and the portion of the world that the ontology is representing. Here, we might ask if the ontology content is correct with respect to the domain it is aiming to represent – does it capture well the structure of the domain and the best of scientific knowledge in that area? Or indeed, we might ask to what extent the ontology is complete with respect to the vocabulary and elements of the domain it aims to cover. A fourth dimension of quality relates to how the ontology is being developed and the community of people it aims to serve. Is it being developed following a community-involving process? Is it openly available? Are its development decisions transparent and well documented? Does it strive to represent and build consensus on the scientific debates at the frontier of its domain, rather than representing a singular and particular perspective, and keep up to date as the scientific frontier advances? Finally, a last dimension of quality relates directly to usability. Here, we can ask whether the ontology is easy to use, and very practically, whether it is usable in one or more applications in such a way as to enhance that application.

In practice, there are various different ways to quantify and thereby concretely measure the quality of an ontology, related to the different intuitive notions of quality that have been described above. For example, some automated tools exist, such as for example OOPS the Ontology Pitfall Scanner (Poveda-Villalón, Gómez-Pérez and Suárez-Figueroa, 2014), that can detect and report on technical and structural infelicities, such as metadata completeness – e.g. whether each class has a unique label and a definition – logical correctness, e.g. whether the ontology overall contains any inconsistencies, and structural whether there are many parent classes that contain only one child class – which would be considered a poor structural feature. Another automated ontology evaluation tool is OntoKeeper (Amith *et al.*, 2019), which in addition to checking syntactic and structural veracity, checks a form of use by virtue of the number of links to that ontology within other ontologies. The widely used ontology software library ROBOT (Jackson *et al.*, 2019) offers a "report" function which runs a series of quality control tests over the input ontology and generates a report file based on the results, suitable for use in an automated workflow.

There are also approaches that aim to quantify coverage of an ontology with respect to an underlying domain vocabulary, often as given by analysis of a corpus of domain-relevant literature. The community-related aspects of ontology evaluation are the focus of the evaluation efforts of the OBO Foundry organization mentioned above (Smith *et al.*, 2007), which has an editorial working group that manually reviews ontologies against criteria such as their openness and community-involvement in their development. Success of an ontology is also often quantified in simple terms of the range and extent to which it is used, which could be represented for example by a number of citations or of posts on a community-wide issue tracker. (By this metric, the Gene Ontology is by far the most successful ontology.) Some of the above criteria for ontology quality are also criteria that underlie how successful an ontology will be, for example, is it easy to use? However, other aspects may drive or preclude wide ontology adoption, including tool support and social and technological `readiness' within the community. Ontologies can also be evaluated in use for specific applications, e.g. the Emotion Ontology was evaluated in use for capturing the self-report of emotions (Hastings *et al.*, 2014). Here, it is clear that evaluation for one particular type of application does not imply appropriateness for a different type of application, but of course, evaluation against multiple types of application provides robust support for the general-purpose quality of the ontology within its domain. Importantly, ontologies are seldom "finished" – rather, they are living and growing entities that need to continuously be maintained. Therefore, it is helpful if quality indicators can be applied and re-applied regularly in order to track progress as the ontology evolves.

Formal Representation and Reasoning with Ontologies

Ontologies are usually represented in computable, logical languages that allow the meanings encoded in them to be made interpretable by computational systems. As such, ontologies are a derivation of early efforts to represent and reason with knowledge computationally, which started in the 1950s and proliferated with the development of rules-based expert systems in the 1970s and

1980s. The knowledge bases that accompanied early expert systems were based on rules of the form *if X, then Y* and were applied in contexts such as medical diagnostic support systems, where the preconditions for the rules would take the form of patient symptoms and laboratory measurements, and the outputs would be differential diagnoses of diseases and conditions.

While rules-based systems are still in use today, in modern computing systems there are a wide range of different formal languages available for the representation of knowledge in ontologies, which reflect different tradeoffs between the desirable objectives of expressivity, which captures how well the language supports encoding the nuances of formal definitions, and automated inference capabilities. In general, the more expressive a language is, the more difficult (and slow) the procedures for automated inference are. Highly expressive logical languages are not decidable in general, which means that no automated approach exists which is able to guarantee determining all the inferences from the expressed logical axioms.

Examples of logical languages that are used for ontologies include variants of first-order predicate logic, such as Common Logic, the Web Ontology Language (OWL), the Open Biomedical Ontologies (OBO) language widely used by ontologies in the biomedical domain, and simpler representational formalisms such as JavaScript Object Notation (JSON). The most widely used of these is OWL, largely due to its adoption as a standard by the W3C for the Semantic Web community, and the wide availability of supporting tools.

Within logic-based languages such as OWL, statements in ontologies have a definite logical meaning within a set-based logical theory, and this is what allows automated tools (known as *reasoners*) to derive inferences and detect errors in ontologies encoded in these languages.

Classes have instances as members, and logical axioms define constraints on class definitions that apply to all class members. For example, the statement "yoga *is a* physical activity" has the logical meaning that all instances of yoga are also instances of physical activity. This can be expressed as the logical axiom *for all x, if x is an instance of yoga, then x is an instance of physical activity:*

$\forall x\colon Yoga(x) \rightarrow PhysicalActivity(x)$

The OWL language is built on top of a family of logical languages that are collectively called Description Logics – in the plural because there are different variants which encompass different logical statement types, and as a result have reasoning algorithms with different levels of complexity. These languages specify logical language elements that allow the formal representation of meanings in the ontology through the definition of axioms. Some of these different ingredients of logical axioms that are available in the OWL language are explained in Table 1 – quantification, cardinality, logical connectives and negation, disjointness and class equivalence.

Language component	Informal meaning	Examples in OWL Manchester Syntax
Quantification: universal (only) or existential (some)	When specifying relationships between classes, it is necessary to specify a constraint on how the relationship should be interpreted: universal quantification ('only') means that for all relationships of that type, the target has to belong to the specified class, while existential quantification ('some') means that at least one member of the target class must participate in a relationship of that type.	*intervention* **has_part** some *intervention delivery* *hydrocarbon* **has_part** some *hydrogen* *hydrocarbon* **has_part** some *carbon* *hydrocarbon* **has_part** only (*hydrogen* or *carbon*)

Cardinality: exact, minimum or maximum	It is possible to specify the number of relationships with a given type and target that a class must participate in, or a minimum or maximum number thereof.	*human* **has_part** exactly 2 *leg*
Logical connectives: intersection (and) or union (or)	It is possible to build complex expressions by joining together expression parts using the standard logical connectives 'and', and 'or'.	*vitamin B* equivalentTo (*thiamin* or *riboflavin* or *niacin* or *pantothenic acid* or *pyridoxine* or *folic acid* or *vitamin B12*)
Negation (not)	In addition to building complex expressions using the logical connectives, it is possible to compose negations.	tailless equivalentTo not (**has_part** some *tail*)
Disjointness of classes	It is possible to specify that classes should not share any members.	*physically active* disjointFrom *sedentary*
Equivalence of classes	It is possible to specify that two classes, or class expressions, are logically equivalent, and that they must by definition therefore share all their members. Logical equivalence is often used to assign *logical definitions* to classes (i.e. necessary and sufficient conditions on class membership), which can be used to automatically infer class placement in a complex hierarchy.	*smoker* equivalentTo (*person* and **participates_in** some *smoking activity*)

Table 1: A selection of logical constructs from the OWL language. Examples are given in OWL's human-readable Manchester syntax (Horridge and Patel-Schneider, 2012).

Like the example axiom given for yoga and physical activity above, each of the axiom examples listed in Table 1 can be expressed as a logical statement in a fragment of the first-order predicate logic. With these axioms, logic-based automated ontology *reasoners* are able to check for errors in an ontology and derive additional inferences from the knowledge that is captured. For example, if a class relation is quantified with 'only' such as the hydrocarbon example given in the table, which can be expressed logically as follows:

$$\forall x \forall y: Hydrocarbon(x) \land hasPart(x,y) \leftrightarrow Hydrogen(y) \lor Carbon(y)$$

and then if a subclass of hydrocarbon in the ontology has a **has_part** relation with a target other than a hydrogen or a carbon (e.g. an oxygen):

$$Hydrocarbon(a) \land hasPart(a,b) \land Oxygen(b)$$

that class will be detected as inconsistent (a contradiction) and flagged as such by the reasoner.

The semantics of logical languages also allow inferences or entailments to be derived automatically from what is captured. Inferences or entailments are additional expressions that follow by the rules of logic from what has been explicitly declared. For example, given the axiom captured above that states that yoga is a subclass of physical activity, if we know that an individual e.g. Bob participates in an instance of yoga, we can infer that Bob also participates in physical activity. This kind of inference is very useful in enabling query answering over large knowledge bases. Moreover, from the axiom that physical activity is disjoint from (share no instances with) being sedentary, we can potentially derive the additional inference that Bob is not sedentary.

Another type of inference that can be computed using the axioms and the knowledge captured in the ontology is to automatically compute the classification hierarchy. For example, given an axiom defining (i.e. specifying an equivalent class expression) organic molecular entities as those molecular entities that contain carbon, and another axiom stating that hydrocarbons contain carbon, the reasoner would be able to infer that hydrocarbon should be subsumed beneath organic molecular entity in the classification hierarchy. This type of inference is particularly beneficial in the management of larger ontologies, as it reduces the amount of work involved in maintenance of large and often multiply polyhierarchical classification systems. Logical definitions for class expressions, in combination with the use of a logical reasoner, can also enable different systems of classification to be used together with the relative arrangement of their hierarchical expressions being computed automatically. For example, in (Chepelev *et al.*, 2012) logical definitions for chemical classes were captured, allowing class definitions from the ChEBI chemical ontology to be used together with class definitions from the MeSH chemical classification, with the resulting composite hierarchy being completely automatically computed, effectively enabling these two knowledge bases to be merged, as illustrated in Figure 3.

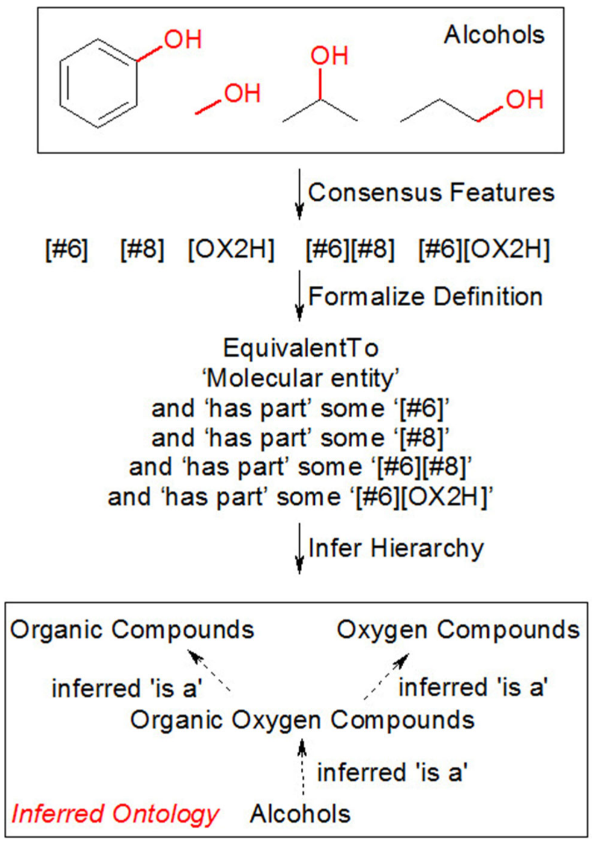

Figure 3: Illustration of logical definition of chemical class features enabling automation of hierarchical arrangement of classes in a chemical ontology, from (Chepelev et al., 2012). The resulting chemical ontology is computable and can be mapped directly to the classification hierarchy present in either ChEBI or MeSH, enabling those diverse ontologies to be automatically merged.

An important distinction about logic-based ontology reasoning when compared to the types of inferences that are commonly used in traditional database systems is that ontology reasoning, at least insofar as it is implemented for the OWL ontology language, assumes what is known as the open world. That means that if an axiom is not captured in the ontology, that does not necessarily mean it is not true. Rather, it means that it is not known whether it is true or not (unless, of course, it can be inferred from other axioms that are captured in the ontology). For example, consider an ontology that describes a class hierarchy of recreational physical activities, and includes classes for yoga, for jogging, and for swimming. If we have an individual Bob who is participating in an instance of recreational physical activity, we can nevertheless not infer that it then must be either yoga, jogging, or swimming. Because the assumption is that the 'world' (or that part of the domain that the ontology represents) is incompletely described, and therefore there might be additional categories of recreational physical activity that are not yet represented. In order to derive an inference of this sort, we would need an additional axiom that captured that all physical activities are either yoga, jogging or swimming. This is sometimes called a covering axiom, or 'closing the world' for a particular part of the domain.

The combination of terminological knowledge with automated deduction, error detection and logical inference capability creates a powerful platform for enabling knowledge-based applications. However, ontologies are not a one-size-fits-all solution, as they have several limitations in practice.

Limitations

Ontologies are just one of a wide range of different knowledge representation structures, each of which has different strengths and weaknesses and is therefore best suited for different purposes and contexts. Modern enterprise knowledge systems typically use a combination of different representational structures and inference components, but there are perennial challenges in interfacing disparate knowledge infrastructures, which presently each such system must address on an ad-hoc and individual basis.

One of the main limitations of ontologies in practice is that, due to their logical basis, they are good at representing statements that are either true or false – known as categorical, or invariant (Rector *et al.*, 2019) – but they cannot elegantly represent knowledge that is vague, statistical or conditional. The types of expressions that are vague in the relevant sense include, for example, the relationships between risk factors and diseases (e.g. eating red meat and cancer), or between socioeconomic status and health outcomes (e.g. poverty and COVID severity). In these examples, there is a statistically significant association at a population level, but it is not the case that every instance of, say, a person living in poverty, is going to be associated with a more severe COVID disease course. Causal models directly represent these types of associative relationships between variables and can be used together with causal reasoning systems that compute over probabilities. Recently, a causal modelling extension of the Gene Ontology was introduced (Thomas *et al.*, 2019) for representing causal knowledge about interacting molecular systems and pathways in combination with the annotations of the functions and activities of genes that are enabled by the standard ontology.

Various extensions do exist at the language level that attempt to combine probabilistic knowledge of various sorts with ontologies. For example, what are called fuzzy logics consist of statements that are not categorical (true or false) but rather are true to some extent, or have vague or probabilistic boundaries (Costa and Laskey, 2006; Lukasiewicz, 2007; Bobillo and Straccia, 2011). Classes that derive their meaning from comparison to a dynamic or conditional group (e.g. the shortest person in the room, which may vary widely) are also not possible to represent well within ontologies. Another form of knowledge that is not well covered in typical ontologies is what is known as default knowledge or defeasible knowledge. For example, the statement that all birds can fly may be true in most cases, and is certainly the case for the exemplar or default examples of what a bird is, but it is not true for all birds. Penguins and ostriches constitute counter-examples. Extensions to the OWL language have also been developed for this type of defeasible knowledge representation (Casini *et al.*, 2015).

Furthermore, despite massive advances in computing technologies and algorithms, there are still pragmatic limits to ensure the scalability of the reasoning tools. For this reason, higher order logical statements, non-binary relationships and other complex logical constructs cannot yet be represented and reasoned with in most of the modern ontology languages. For an example of the practical implications of this restriction, it can be difficult to adequately capture knowledge about change over time at the class level, i.e. classes in which the members participate in relationships at one time and not at another, as including a temporal index for each relationship assertion would require ternary relations: the statement being expressed is not just of the form *x* is-related-to *y*, which is a binary statement as there are two variables *x* and *y*, but rather of the form *x* is-related-to *y* at *t*, which is a ternary statement, as there are three variables *x*, *y* and *t*. The standard form of the OWL language does not support relationships with an arity (number of entities related) of higher than two, although again, there are extensions to address this requirement (e.g. Salguero, Delgado and Araque, 2009).

In practice, none of the OWL extensions surveyed in this section are in widespread use, and indeed the largest portion of ontologies in contexts such as the semantic web are still rather inexpressive. It is clear that the time taken for a reasoner to perform its computations is still a barrier in practice for many large-scale data-driven use cases, and indeed it is the most lightweight profile (sub-language) of the OWL language, known as EL, for which there are very rapid, parallel and incremental reasoning algorithms available (Kazakov, Krötzsch and Simancik, 2012), that is the most widely used in practice.

Machine Learning with Ontologies

Machine learning is a family of approaches falling within the broader family of artificial intelligence technologies in which computers 'learn' to make complex predictions based on data, facts or text. The performance of such predictions is then evaluated by comparison of the predictions to the truth in a dataset for which the expected outcomes are known for a set of inputs.

A range of different approaches exist to enable machines to learn, including some that are logic-based, some that are based on mathematical equations and some that are based on statistical approaches and data. However, many of the breakthroughs of modern artificial intelligence technology in real-world applications, such as for example recognizing people or objects in images, have been due to what are called 'deep' learning artificial neural networks (Figure 4). The performance breakthroughs of such networks regularly feature in the news – for example, a recent major breakthrough in the prediction of protein three-dimensional structures from amino acid sequences was achieved using an architecture based on deep neural networks (Jumper *et al.*, 2021).

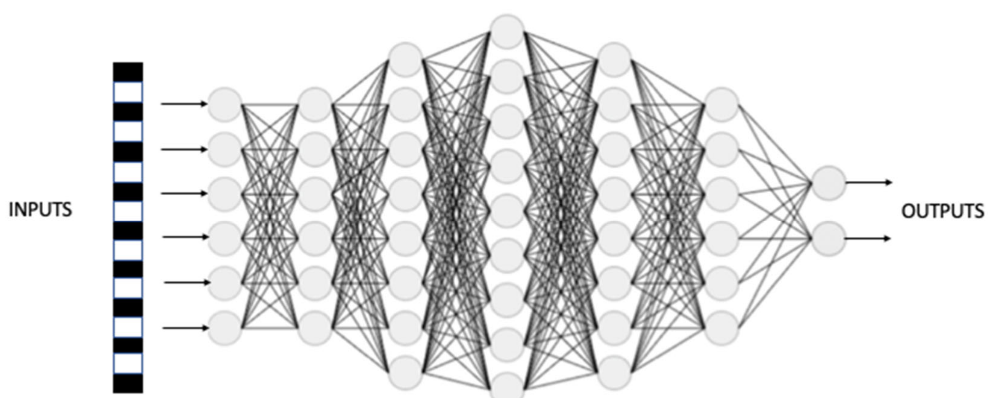

Figure 4: Deep neural network architectures consist of many layers of interconnected nodes.

Deep neural networks contain many layers of interconnected nodes that are each individually able to integrate inputs and form outputs, that then go on to serve as the inputs to the next layer, each of which having their own trainable parameters and state. Modern deep neural network architectures are large and complex structures that allow information to flow and be synthesized in both directions. Information flows forward from inputs to outputs, as each node in the network accepts several inputs and provides an output to the nodes in the next layer. However, information also flows backward in the form of "back-propagation", i.e. information about the error in the final prediction, allowing information about how well the network has performed in some training task where the correct output is known to be used to provide input to the internal parameter settings of the nodes in a way that improves their performance on subsequent iterations of the training process. This type of network learns an overall function with a vast number of parameters – each node and connection in the network having its own parameters – which are typically set during the training process using a gradient descent optimization algorithm that is optimizing towards the maximally successful prediction output. For that reason, because there are so many parameters that have to be set for the network to be able to make good predictions, deep neural networks are only feasible when there are large quantities of training data available to use for training the internal state of the network.

Research into deep neural network architectures is ongoing and very prolific. Novel elements are regularly added to such frameworks and where such elements bring a significant improvement on the state of the art, those architectures tend to become the more commonly used. For example, just a few years ago a mechanism was introduced that allowed for a selective across-network *attention* state to be maintained at each layer of a deep neural network, in addition to the parameters each node maintains which synthesize inputs into outputs (Vaswani *et al.*, 2017). Since then, networks using attention mechanisms have become very widely used and outperform earlier approaches for a multitude of applications.

As representations of domain knowledge and thereby specifications of the 'ground truth' in a particular domain, ontologies are often part of the input data to machine learning systems. Such systems typically learn from large quantities of annotated data – that is, data to which labels from a structured controlled vocabulary have been assigned. Ontologies provide controlled vocabularies that may be used for the structured annotation of the text and data which are then the inputs to learning models. In biomedical natural language processing, for example, one of the most frequent tasks that machine learning is used for is the identification of novel entities of a given type in a corpus of text (Song *et al.*, 2021) – for example the identification of genes and their relationship to scientific interests and activities (Serrano Nájera, Narganes Carlón and Crowther, 2021) or the identification of drugs and their interactions (Herrero-Zazo *et al.*, 2015). In most such methods, the gold standard that is used for training the networks that subsequently make predictions consists of a corpus of text annotated using the vocabulary from one or more biomedical ontologies – for example, abstracts of text in which the words or phrases which represent diseases have been tagged with the semantic concept "disease" and associated with an identifier from a disease ontology. In the behavioral sciences, a system that is able to make predictions about the effects of interventions based on synthesis of all the available evidence about behavioral interventions (Michie, Thomas, *et al.*, 2020) is trained from a corpus of scientific papers annotated using terms drawn from an extensive ontology for behaviour change interventions (Michie, West, *et al.*, 2020), which specifies the structure of interventions, their content and their parameters.

However, approaches that make use of annotated text often do not use the *semantic* content of the ontologies at all – they do not use the logical axioms, nor the textual definitions, nor (with some exceptions) even the relationships between entities. Rather, they use ontologies only as controlled vocabularies which, intuitively, one can view as a "set of buckets" into which content is sorted and with which content is tagged. This is a shame, because the knowledge encoded in the ontologies has the potential to substantially enhance the performance of common tasks such as recognizing entities of a given type or predicting relationships between entities. Approaches which make more sophisticated use of the content of ontologies require that the ontology content itself be made amenable to serve as inputs to learning algorithms directly, which requires an appropriate transformation of the ontology content into a mathematical form.

Semantic Embeddings

Machine learning algorithms operate by finding patterns or regularities in their inputs that are characteristically associated with the outputs. The inputs to such algorithms need to first be transformed into a suitable mathematical format for the pattern learning process to work – usually vectors, or long series of numbers. A significant part of the effort in creating a neural network architecture to address a given problem is solving the challenge of how to transform the inputs into a relevant format. One straightforward strategy is to assign a "one hot" encoding for the input content, that is, having a vector of the same length as the number of different values that the input can take, and then for a given input instance, specifying '1' if that value is present and '0' otherwise. The equivalent for natural language is to assign each word a position in a vector, and encoding sentences based on which words are present in the sentence. Clearly, for the most part such vectors will be large and information-poor (sparse), since some words will appear only very infrequently, and the words that appear most frequently will convey the least information. Thus, various more sophisticated algorithms exist to derive encodings for text and data.

More recent neural network architectures use approaches which are able to determine the best representation for a given raw input directly by learning from a large set of data, based on inferring an optimal vector to represent a given portion of the input from the context of the input. Such trained representations known as *embeddings,* and have the favorable property that similar words have similar embeddings, as they are used in similar contexts. In natural language processing, for example, huge corpora of text are used as the input to learn embeddings for each word by training a network with the task of predicting each word from its context (the words around it in the corpus). These learned word embeddings, known as language models, are then available for re-use in other tasks. An example of a very widely used language model is BERT (Devlin *et al.*, 2019).

To use the semantic content of ontologies to enhance machine learning predictions, the ontology content itself needs to be encoded or embedded into a suitable form to serve as input to a neural network. Approaches to embedding ontologies for machine learning can be grouped into three main categories (Kulmanov *et al.*, 2020). Firstly, graph-based approaches treat ontologies as graphs and draw from algorithms for embedding graph-like structures into vectors. In these approaches, each node in the graph is taken to represent a unit of input, and the context for the node is given by the immediate neighborhood of the graph – the node's edges, the nodes that it is connected to, their edges in turn, and so on. Ontologies captured in the OWL language are not inherently graph-structured, but rather they are structured in the form of axioms with logical meanings. Thus, graph-based embedding approaches first need to determine a representation of the ontology content as a graph, for example, by transformation to an RDF graph using the standard RDF semantics for the OWL language. Thereafter, random walks through the graph structure are often used to generate the graph-based embeddings, which may be weighted or biased to emphasize certain features over others. For example, OWL2Vec (Holter *et al.*, no date) is an embedding approach that uses random walks around class nodes in a graph generated from the ontology to create embeddings, which was further extended with annotation information (such as labels, synonyms, comments and textual definitions) in OWL2Vec* (Chen *et al.*, 2020). Secondly, syntactic approaches treat ontology axioms as text similar to sentences and aim to embed this content in a way that preserves syntactic regularities (such as frequencies of co-occurrences). Examples of syntactic approaches to embedding ontologies include OPA2Vec (Smaili, Gao and Hoehndorf, 2019), which first transforms logical axioms into sentences and then uses a sentence-based embedding approach to create vectors in a similar way to how natural language texts are processed. Thirdly, there are embedding approaches that aim to preserve the logical semantics of the content using model-theoretic properties. For example, these approaches may draw on grounded real-valued logics to convert logical axioms directly into vectors in such a way that the logical properties can be preserved in vector operations.

Beyond ontologies 'proper', there are also many applications of modern graph-based machine learning approaches to process, extend and learn from what are called 'knowledge graphs' – composites of the simpler, less expressive elements of ontologies, together with large-scale data, structured into vast semantic networks. Knowledge graphs are the family of technologies that underlie most modern search engines and recommendation systems, such as those of tech giants Google, Facebook and Amazon. An important application of machine learning with knowledge graphs is what is known as "link prediction" or "knowledge completion", i.e. predicting new relationships between entities that are in the graph but not yet related.

Machine learning with semantic information arising from ontologies has been applied to several challenging knowledge integration and synthesis tasks in the biomedical domain, including the prediction of associations between drugs and their target proteins (Alshahrani *et al.*, 2017), to interpret and visualize the phenotypic diversity associated with genomic variation in mice (Konopka, Vestito and Smedley, 2021), to predict the associations of genes with diseases (Nunes, Sousa and Pesquita, 2021), and in the behavioral sciences to train a system to make personalized exercise recommendations (Lv *et al.*, 2018). Although such approaches are relatively recent, there are already good indications that such methods outperform their purely data- or corpus-based predecessors.

Limitations

Machine learning approaches in general suffer from a number of known limitations (Pearl, 2018; Marcus, 2020). They are known to susceptible to easily being misled, a phenomenon that is called 'brittleness', for example, in which certain small alterations of the input sequence can lead to completely different output predictions. They lack explainability, that is, they do not give explanations or justifications for their predictions and therefore may be difficult to trust. They are very data- and compute-resource intensive, requiring far, far more data to learn from than typical humans would need to learn similar things, and leading to concerns about environmental sustainability. And, while they may show compelling performance when being developed and evaluated in controlled scenarios with the use of metrics designed to measure their performance in making correct predictions against a defined gold standard, translation of such systems into real-world applications is often challenging (Futoma *et al.*, 2020). Subtle biases and assumptions pervade the test data that is used in the algorithmic training, and then when the system is deployed on real-world data the performance of the system is significantly poorer (Kelly *et al.*, 2019).

Moreover, such models may learn human biases that are present in their training data, such as prejudices against women or persons from minority groups (Bender *et al.*, 2021). Many of these types of prejudices can be seen to arise naturally from the combination of unbalanced input data, combined with a lack of *understanding* on the part of the model. For example, even in modern machine translation models, which perform on average very well for many translation tasks, reliably translate the English 'they' to the French 'ils' (the masculine form of they), even in a sentence such as 'They gave birth to their firstborn children.' (Translated as *'**Ils** ont donné naissance à leurs* *premiers enfants'* by DeepL). Machine learning models only contain *statistical, associative* knowledge, and 'ils' is far more common in French language expressions that translate 'they' than 'elles' is. Ontologies, on the other hand, contain axioms that specify constraints that must always apply, such as that – canonically at least, for humans, only females can give birth. But this type of knowledge is not usually accessible to neural-network-based learning systems. And at present, even most approaches that combine semantic information from ontologies into neural networks only include *associative* semantic information. That is, they can be used to enhance a learning process by encompassing the network neighborhood of a class in an ontology, specifying which classes are closely related – or by representing which classes are more similar in terms of sharing similar relationships or axiomatic patterns. But they cannot be used to encode constraints which must be adhered to in an axiomatic, rules-based fashion, such as that men do not give birth. The standard architectures of modern neural networks simply do not allow for logical expressions, rules or constraints, nor for reasoning, consistency checking or inference.

There are emerging novel architectures that aim to address this limitation. For example, Logical Neural Networks (Riegel *et al.*, 2020) offer an architecture for a neural network in which each neuron has a defined logical meaning as a component of a logical formula. That is, a neuron can represent an *and* operation, or an *or* operation, or negation. The network can then be trained on data and will try to learn a representation that is maximally consistent with the overall logical formula that is expressed by the combinations of the configured neurons. This is one example of a "neuro-symbolic" (Garcez and Lamb, 2020) approach to neural network architecture, which aims to combine elements from knowledge representation and symbolic reasoning (such as is used in ontologies) and statistical learning (such as is used in more typical neural networks). On their own, such logic-driven networks may be of perhaps limited use, in their being able to do certain forms of logical reasoning that could also be done symbolically, but their great potential lies in the fact that they can be combined with other networks in layered, hybrid architectures which are then able to combine traditional associative learning with the use of logical inference or the enforcement of certain sorts of logical constraints, such as have been used in practice for example to develop a sophisticated query answering system over public knowledge graphs including WikiData (Abdelaziz *et al.*, 2021). It is widely anticipated that such hybrid systems will become more and more essential in the future to address the limitations of the current generation of deep learning systems, leading some to say "if the aim is to build a rich AI system, that is, a semantically sound, explainable and ultimately trustworthy AI system, one needs to include with it a sound reasoning layer in combination with deep learning" (Garcez and Lamb, 2020). The article (Garcez and Lamb, 2020) sets out a vision for a "third wave" of AI systems in which symbols and symbolic reasoning act as constraints on sub-symbolic networks and thereby help to improve learning performance as a part of a continuous positive cycle of feedback between learning and reasoning – a vision which the use of ontologies together with neural networks is well-positioned to turn into a reality with immense benefits across scientific domains.

Automatically Extending Ontologies

One of the main limitations of the broad use and adoption of ontologies and other formal knowledge representation structures is the time it takes to capture and encode content into them. This is known as the 'knowledge acquisition bottleneck'. Typically, ontologies capture expert knowledge, and require ample time from experts – experts who are familiar with the logical structures of the language – in their development. As such, there is great interest in algorithms and approaches that are able to automate the assembly of content into ontologies, or those that are able to extend existing ontologies with novel content.

Early attempts to automate the acquisition of knowledge into ontologies arose out of the methods of natural language processing, that is, automated approaches to interpreting and parsing text captured in a natural (human) language, such as English or French. In most ontology learning approaches (Gómez-Pérez and Manzano-Macho, 2003; Asim *et al.*, 2018) natural language processing is used to process the textual data with the objective of identifying terms for entities that are important for the domain and that should correspond to classes in the ontology, together with their relationships, including hierarchy and other semantic relationships. Simple approaches identify noun phrases using linguistic processing, and use simple rules-based patterns to extract relationships between these noun phrases. More sophisticated approaches use corpus-based contextual information, such as the words that co-occur the most frequently, or in modern language models the learned contextual embeddings, to derive meanings for and thereby semantic distances between words or phrases such that they can be clustered or aggregated.

A widely used ontology learning from text approach, Text2Onto (Cimiano and Völker, 2005), uses a combination of algorithms to suggest novel ontology content. From a corpus of text, it initially performs linguistic processing in order to determine parts of speech and extract words or phrases that are anticipated to relate to classes and to relations. This part of the overall architecture is rules-based and language-specific. The default available different implementation is the English language, with separately available modules supporting other languages such as, for example, Spanish (Völker, Fernández and Sure-Vetter, 2008) and more recently French (Hajji, Qbadou and Mansouri, 2020). The candidate novel classes that are linguistically extracted are then subjected to several quantitative algorithms to predict their relevance for an ontology, including relative term frequency (the number of times term appears in a document relative to the total number of terms in the document), term frequency – inverse document frequency (down-weighting terms that appear in nearly all documents), and a commonly used method called C-value / NC-value that aims to separate meaningful multi-word terms from their surrounding contexts (Frantzi, Ananiadou and Mima, 2000). Each of these methods give a quantitative value that is associated with the relevant term and then normalized for use as a probability which is associated with the prediction in the Text2Onto tool. Having quantitative probabilities allows for user-defined thresholds which allow custom decisions about filtering between predictions and noise. The tool also contains various algorithms for relation extraction from the text, which in turn are also associated with probabilistic scores. Finally, the tool presents the highest-ranked predictions to the user via a graphical interface, which allows the user to select and tune the resulting ontology by adding or removing classes or relations, before saving the resulting new ontology.

While ontology learning focuses on creating new ontologies, in scientific contexts where there is often a pre-existing ontology, the extension of a pre-existing ontology with novel content is an even more important scenario. Approaches to ontology extension are broadly modelled on those of ontology learning, while often using the content of ontology as a ‘seed’ to identify additional terms and phrases that are important for the target domain. For example, in (Liu *et al.*, 2005) a system is proposed that uses co-occurrence analysis between seed terms and discovered terms in a corpus to rank candidate terms for inclusion in an extended version of an ontology. Term discovery is followed by term disambiguation by reference to a dictionary, and rules-based determination of subsumption (hierarchical) relationships from the corpus to connect novel terms to the original seed ontology. An alternative paradigm uses approaches from machine learning and statistics to suggest novel content for inclusion, for example based on text clustering (Liu and Li, 2018) or topic models. An exemplar of such approaches is the recent one introduced in (Li, Armiento and Lambrix, 2019) that builds on phrase-based topic models (El-Kishky *et al.*, 2014) to extract phrases with close relevance to the topics given by the ontology that is being extended, and then uses a 'concept lattice' approach on the resulting phrases in order to rank and structure candidate phrases for inclusion in the ontology.

It should be noted that automatically assembled ontology extensions are commonly rather noisy, and may potentially introduce biases or errors from the literature into the ontology. In general, all approaches which aim to extract conceptualizations from text corpora are subject to high levels of noise. To mitigate against such problems, many of the automated ontology extension approaches involve several manual steps, in which experts evaluate concepts and related phrases to sort out these potential issues. For example, the topic-modelling approach described above which has been applied to extend ontologies in the materials science domain (Li, Armiento and Lambrix, 2019) makes extensive use of human selection in interpreting the results and deciding which of the suggested and ranked content entities should be included in the resulting ontology. The examples of decisions which the pipeline is not able to automatically make is whether the extracted topic is too general for the stated domain (e.g. 'electron transfer' for a materials science domain ontology), or is an exact synonym of an entity that is already included in the ontology, or in fact is a novel entity that should therefore be included. Of course, there are also examples of candidates that are extracted and yet are not relevant for the ontology at all (noise). Moreover, even if candidate phrases are good recommendations for inclusion in an extended version of the ontology, the existing approaches do not retrieve all of the required metadata, such as definitions, for these added classes. The need for manual intervention suggests that for the foreseeable future, "human-in-the-loop" approaches may still dominate knowledge acquisition rather than fully automated approaches.

The approaches to extending ontologies based on clustering and contextual similarities between known ontology content and phrases in a corpus are clearly related to the similarity between words and phrases that is represented by their learned embeddings in modern language models. Thus, modern language models in combination with pipelines that are able to combine machine learning with ontology content seem to offer a promising technology for significantly improving ontology extension approaches. And indeed, there are some promising initial efforts in this direction. In a recent, state-of-the-art implementation of ontology extension through learning (Althubaiti *et al.*, 2020), the target domain is diseases and the target ontology is the Disease Ontology (DO). The implementation starts by using an automated dictionary-based text annotation tool (WhatIzIt) to automatically annotate a relevant corpus of text. They then use Word2Vec (Mikolov *et al.*, 2013) to generate contextual embedding vectors for the corpus. These embeddings are subsequently used to suggest words or phrases that are similar enough to the existing ontology content to count as a synonym, subclass or related class of one of the classes in the ontology. To predict which of these novel candidates are subclasses of the root class 'disease' rather than merely related classes, they train an artificial neural network classifier to make this prediction. They evaluate classifiers able to distinguish infectious from anatomical diseases (a fundamental distinction in the Disease Ontology). They further evaluate extending this distinguishing capability to several of the sub-classes of these two classes in a multi-class classification approach. However, the performance of the neural network is determined to decrease rapidly with the number of classes added to the multi-class classification task, which in the end still only tackles fewer than 10 class distinctions, far short of the overall number of classes in the disease ontology classification. Moreover, the whole pipeline still only operates within one semantic domain – custom classifiers would need to be trained for each semantic domain for which this approach needs to be applied.

We might envision that some of the hybrid approaches to machine learning discussed above in the section on machine learning with ontologies might have applicability also to the problem of ontology extension. For example, noise might be reduced by using constraints applied to a neural network through, for example, including a logical neural network layer in the overall architecture. This would allow domain experts to specify constraints on the type of recommendations that are returned by the system in a way that reduces the overall search space and may make the resulting recommendations have higher quality. However, such architectures do not appear to have been applied to this problem in practice yet.

Applications in the Behavioral Sciences

Historically, ontologies have been adopted at scale mainly within the biomedical sciences. However, in recent years they have started to see adoption more widely within other scientific domains such as ecology, agriculture, economics, and the behavioral sciences. The Human Behaviour-Change Project (HBCP; Michie, Thomas, *et al.*, 2020), mentioned above, has developed a large-scale suite of ontologies for the domain of human behavior and behavioral interventions, which has been applied to organize and mine the literature using a combination of manual annotation and machine learning. Automation within the HBCP consists of two separate components – an information extraction component, which aims to identify all the relevant aspects of interventions within the full texts of intervention reports, and a prediction component, which learns from both manually annotated and automatically extracted intervention reports to predict the effectiveness of interventions given a combination of parameters describing the intervention scenario. The outcome prediction model of the HBCP consists of a deep learning neural network model with a combination of ontology-based annotations and text as inputs, thus, it exemplifies the use of ontologies together with machine learning to solve a challenging task in the automation of analyses.

The HBCP project has covered new ground in several different ways and has also been able to derive several learnings about the particular challenges that arise when applying these types of technologies to the behavioral sciences. Firstly, we can note that in contrast to the biomedical domain, the terminology that is used in scientific studies in the behavioral sciences overlaps with informal and colloquial usages of vocabulary and terminology across multiple other domains. While in the biomedical sciences, technical vocabulary includes gene names, drug names and disease names, each of which is strongly semantically typed and forms a vocabulary that for the most part is distinct from the vocabularies used in other sciences and domains, in the behavioral and social sciences the terminology that is used reflects human activities and concerns, social structures and patterns of behaviour. These entities are also referred to frequently within other disciplines, as many studies that involves humans, even biomedical studies of drugs and diseases, will ultimately need to refer to population groups, activities and behaviors. Moreover, a good portion of the vocabulary used in the behavioral and social sciences have colloquial meanings, such as for example when the term ‘behavior’ is used in expressions such as “cancer genes may behave distinctly in different experimental settings” or “the hippocampus behaved like other DMN regions”. Here, the meaning of the term ‘behavior’ is very different from the technical sense in which it is used in the study of human behavior, which is clearly one of the core semantic categories within the behavioral and social sciences. The learning of embeddings from contexts of use in large-scale corpora of text – language modelling – can be expected to perform more poorly on the specific technical uses of such terminology when there are a very large number of colloquial uses outnumbering the technical uses. Moreover, the same terminology may have very different implications or meanings in different contexts. For example, quitting in a smoking cessation intervention is the desirable objective, while in a physical activity intervention it is an undesirable outcome or side effect.

Relatedly, there is a lack of an organized repository for the literature relevant to the behavioral sciences. In the biomedical sciences, organized repositories such as PubMed exist and can be taken as the definitive source for relevant literature for the domain, comprehensive with respect to abstracts, and even encompassing a substantial portion of open-access full text. A great number of the semantic text-based methods that have been developed for the biomedical domain depend on the existence of open and accessible repositories of relevant text and data. In the behavioral sciences, in contrast, the relevant literature spans several different disciplines including psychology, the social sciences, and economics as well as the more medically-oriented branches e.g. behavioral medicine. These disciplines may have quite different publication practices to those in the biomedical domain, and in some of these even discovering the existence of the relevant literature may be subject to a paywall, e.g. psychology (JSTOR). Moreover, the challenge with acquiring content for training learning systems is not limited to the availability of textual data in the form of scientific publications. There is also a much larger variety of data available in the public domain in the biomedical sciences than in the behavioral sciences, including for example structured and ontology-annotated data such as the associations of genes to functions, or databases of drugs and metabolites and their names and synonyms. In the behavioral sciences, there is significantly less publicly available data, and any research data involving humans is subject to privacy and ethical concerns that may mean that this situation will not change dramatically in the future.

Finally, the behavioral science domain is characterized by a very large number of distinct semantic types. In biomedical contexts, it is more common that although there may be a large number of named entities within a given domain or application area, there are a relatively small number of semantic types (e.g. drugs, biological processes, diseases). This means that automated identification of new entities belonging to a given semantic type can benefit from large numbers of training examples, as there are many examples of existing entities belonging to that semantic type. However, in the behavioral sciences, there are many different semantic types – individuals, populations, interventions, the full range and scope of behaviors, population and personal attributes, location attributes, and moreover all the semantic types from biomedicine which may also have relevance in some intervention contexts. Added to this semantic complexity, there are very heterogeneous ways in which such semantic types are described in the literature, for example, appearing in text descriptions or in tables. Moreover, the theoretical frameworks in which such research is conducted may introduce technical terminology from different theoretical backgrounds which may be difficult for a language model to disentangle, as there are a wide range of different theories and constructs in use within behavioral science more broadly, and no systematic way to map their overlaps and commonalities.

Taken together, these observations suggest that machine learning methods will perform more poorly in the behavioral domain when compared to their performance in the biomedical domain. Nevertheless, the semantic clarity afforded by community-wide efforts to develop ontologies to survey, map, structure and organize the content of the behavioral domain will help, as will concordant efforts to open up the associated literature and to increase the scope of available data. In addition, modern hybrid approaches which are able to harness a diversity of sources of inputs – ontologies, data, text corpora – and harness them for learning as well as logic based reasoning and causal or theoretical inferences – will increase the power of both learning and reasoning approaches by allowing them to be used in combinations that harness the strengths of each.

Conclusions

Reasoning about the full complexity of human behavior in context may be one of the most complex tasks which we can set ourselves – and as such, it is no surprise that it is also difficult to teach to machines. Nevertheless, many advances have been made recently towards formalization in this domain, and it is exciting that this coincides with a time during which we are witnessing a potential technological explosion of new methods that are able to harness formalizations such as ontologies together with powerful data-driven learning algorithms. There is much work still to do to bring the wealth of knowledge and expertise in the behavioral domain into computable formats and to make sufficient structured data and text available for large-scale learning algorithms to consume. The most exciting frontiers lie at the interfaces between knowledge and inference, between learning and reasoning, between meanings and patterns. To this end, it can even be anticipated that the behavioral sciences can have a positive impact on the computational sciences, both by providing challenging use-cases for the development of novel methods, and also through synthesizing knowledge about human reasoning and behaviour, which in turn can inform the development of artificial approaches.

References

Abdelaziz, I. *et al.* (2021) 'A Semantic Parsing and Reasoning-Based Approach to Knowledge Base Question Answering', *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(18), pp. 15985–15987.

Alshahrani, M. *et al.* (2017) 'Neuro-symbolic representation learning on biological knowledge graphs', *Bioinformatics*. Edited by J. Kelso, 33(17), pp. 2723–2730. doi: 10.1093/bioinformatics/btx275.

Althubaiti, S. *et al.* (2020) 'Combining lexical and context features for automatic ontology extension', *Journal of biomedical semantics*, 11(1), pp. 1–13.

Amith, M. *et al.* (2019) 'Architecture and usability of OntoKeeper, an ontology evaluation tool', *BMC Medical Informatics and Decision Making*, 19(4), p. 152. doi: 10.1186/s12911-019-0859-z.

Arp, R., Smith, B. and Spear, A. D. (2015) *Building Ontologies with Basic Formal Ontology*. The MIT Press.

Ashburner, M. *et al.* (2000) 'Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.', *Nature Genetics*, 25(1), pp. 25–29. doi: 10.1038/75556.

Asim, M. N. *et al.* (2018) 'A survey of ontology learning techniques and applications', *Database*, 2018(bay101). doi: 10.1093/database/bay101.

Bauer, S. (2017) 'Gene-Category Analysis', in Dessimoz, C. and Škunca, N. (eds) *The Gene Ontology Handbook*. New York, NY: Springer New York, pp. 175–188. doi: 10.1007/978-1-4939-3743-1_13.

Bender, E. M. *et al.* (2021) 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? □', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event Canada: ACM, pp. 610–623. doi: 10.1145/3442188.3445922.

Berthelon, F. and Sander, P. (2013) 'Emotion ontology for context awareness', in *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)*. *2013 IEEE 4th*

International Conference on Cognitive Infocommunications (CogInfoCom), Budapest, Hungary: IEEE, pp. 59–64. doi: 10.1109/CogInfoCom.2013.6719313.

Bobillo, F. and Straccia, U. (2011) 'Fuzzy ontology representation using OWL 2', *International Journal of Approximate Reasoning*, 52(7), pp. 1073–1094. doi: 10.1016/j.ijar.2011.05.003.

Brank, J., Grobelnik, M. and Mladenic, D. (2005) 'A survey of ontology evaluation techniques', in *Proceedings of SIKDD 2005*, p. .

Casini, G. *et al.* (2015) 'Introducing Defeasibility into OWL Ontologies', in Arenas, M. et al. (eds) *The Semantic Web - ISWC 2015*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 409–426. doi: 10.1007/978-3-319-25010-6_27.

Chen, J. *et al.* (2020) 'OWL2Vec*: Embedding of OWL Ontologies', *arXiv:2009.14654 [cs]*. Available at: http://arxiv.org/abs/2009.14654 (Accessed: 8 October 2020).

Chepelev, L. L. *et al.* (2012) 'Self-organizing ontology of biochemically relevant small molecules', *BMC Bioinformatics*, 13(1), p. 3. doi: 10.1186/1471-2105-13-3.

Cimiano, P. and Völker, J. (2005) 'Text2Onto', in Montoyo, A., Muńoz, R., and Métais, E. (eds) *Natural Language Processing and Information Systems*. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science), pp. 227–238. doi: 10.1007/11428817_21.

Costa, P. C. G. da and Laskey, K. B. (2006) 'PR-OWL: A Framework for Probabilistic Ontologies', in *International Conference on Formal Ontology in Information Systems*, pp. 237–249.

Davis, E. and Marcus, G. (2015) 'Commonsense reasoning and commonsense knowledge in artificial intelligence', *Communications of the ACM*, 58(9), pp. 92–103. doi: 10.1145/2701413.

Devlin, J. *et al.* (2019) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', *arXiv:1810.04805 [cs]*. Available at: http://arxiv.org/abs/1810.04805 (Accessed: 4 September 2021).

El-Kishky, A. *et al.* (2014) 'Scalable topical phrase mining from text corpora', *Proceedings of the VLDB Endowment*, 8(3), pp. 305–316. doi: 10.14778/2735508.2735519.

Frantzi, K., Ananiadou, S. and Mima, H. (2000) 'Automatic recognition of multi-word terms:. the C-value/NC-value method', *International Journal on Digital Libraries*, 3(2), pp. 115–130. doi: 10.1007/s007999900023.

Futoma, J. *et al.* (2020) 'The myth of generalisability in clinical research and machine learning in health care', *The Lancet Digital Health*, 2(9), pp. e489–e492. doi: 10.1016/S2589-7500(20)30186-2.

Garcez, A. d'Avila and Lamb, L. C. (2020) 'Neurosymbolic AI: The 3rd Wave', *arXiv:2012.05876 [cs]*. Available at: http://arxiv.org/abs/2012.05876 (Accessed: 5 September 2021).

Gómez-Pérez, A. and Manzano-Macho, D. (2003) *A survey of ontology learning methods and techniques*, p. 86.

Hajji, M., Qbadou, M. and Mansouri, K. (2020) 'An adaptation of Text2Onto for supporting the French language', *International Journal of Electrical and Computer Engineering (IJECE)*, 10(4), pp. 3743–3750. doi: 10.11591/ijece.v10i4.pp3743-3750.

Hastings, J. *et al.* (2011) ‘Dispositions and processes in the Emotion Ontology’, in *Proceedings of the International Conference on Biomedical Ontology (ICBO2011), Buffalo, USA*.

Hastings, J. *et al.* (2014) ‘Evaluating the Emotion Ontology through use in the self-reporting of emotional responses at an academic conference’, *Journal of Biomedical Semantics*, 5(1), p. 38. doi: 10.1186/2041-1480-5-38.

Hastings, J. (2017) ‘Primer on Ontologies’, in Dessimoz, C. and Škunca, N. (eds) *The Gene Ontology Handbook*. New York, NY: Springer New York (Methods in Molecular Biology), pp. 3–13. doi: 10.1007/978-1-4939-3743-1_1.

Hayes, P. J. (1989) ‘Naive physics I: ontology for liquids’, in *Readings in qualitative reasoning about physical systems*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 484–502.

Herrero-Zazo, M. *et al.* (2015) *Application of Domain Ontologies to Natural Language Processing: A Case Study for Drug-Drug Interactions*, *International Journal of Information Retrieval Research (IJIRR)*. Available at: www.igi-global.com/article/application-of-domain-ontologies-to-natural-language-processing/132500 (Accessed: 17 November 2020).

Holter, O. M. *et al.* (no date) ‘Embedding OWL Ontologies with OWL2Vec’, p. 4.

Horridge, M. and Patel-Schneider, P. F. (2012) *OWL 2 Web Ontology Language Manchester Syntax*. Available at: http://www.w3.org/TR/owl2-manchester-syntax/.

Horrocks, I. *et al.* (2007) ‘OWL: a Description Logic Based Ontology Language for the Semantic Web’, in *The Description Logic Handbook*. 2nd edn. Cambridge, UK: Cambridge University Press.

Jackson, R. C. *et al.* (2019) ‘ROBOT: A Tool for Automating Ontology Workflows’, *BMC Bioinformatics*, 20(1), p. 407. doi: 10.1186/s12859-019-3002-3.

Jumper, J. *et al.* (2021) ‘Highly accurate protein structure prediction with AlphaFold’, *Nature*, pp. 1–11. doi: 10.1038/s41586-021-03819-2.

Kazakov, Y., Krötzsch, M. and Simancik, F. (2012) ‘ELK Reasoner: Architecture and Evaluation’, in *ORE*.

Kelly, C. J. *et al.* (2019) ‘Key challenges for delivering clinical impact with artificial intelligence’, *BMC Medicine*, 17(1), p. 195. doi: 10.1186/s12916-019-1426-2.

Konopka, T., Vestito, L. and Smedley, D. (2021) ‘Dimensional reduction of phenotypes from 53,000 mouse models reveals a diverse landscape of gene function’, *bioRxiv*, p. 2021.06.10.447851. doi: 10.1101/2021.06.10.447851.

Kulmanov, M. *et al.* (2020) ‘Semantic similarity and machine learning with ontologies’, *Briefings in Bioinformatics*, (bbaa199). doi: 10.1093/bib/bbaa199.

Li, H., Armiento, R. and Lambrix, P. (2019) ‘A method for extending ontologies with application to the materials science domain’, *Data Science Journal*, 18(1), pp. 1–21.

Liu, F. and Li, G. (2018) ‘The Extension of Domain Ontology Based on Text Clustering’, in *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*. *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, pp. 301–304. doi: 10.1109/IHMSC.2018.00076.

Liu, W. *et al.* (2005) 'Semi-automatic ontology extension using spreading activation', *Journal of Universal Knowledge Management*, (1), pp. 50–58.

Lukasiewicz, T. (2007) 'Probabilistic Description Logics for the Semantic Web', *TU Vienna*, infsys research report(1843-06–05).

Lv, P. *et al.* (2018) 'Utilizing knowledge graph and student testing behavior data for personalized exercise recommendation', in *Proceedings of ACM Turing Celebration Conference - China*. *TURC 2018: ACM Turing Celebration Conference - China*, Shanghai China: ACM, pp. 53–59. doi: 10.1145/3210713.3210728.

Marcus, G. (2020) 'The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence'. Available at: https://arxiv.org/vc/arxiv/papers/2002/2002.06177v1.pdf.

McCarthy, J. and Hayes, P. J. (1969) 'Some Philosophical Problems from the Standpoint of Artificial Intelligence', in Meltzer, B. and Michie, D. (eds) *Machine Intelligence 4*. Edinburgh University Press, pp. 463–502.

Michie, S., West, R., *et al.* (2020) 'Representation of behaviour change interventions and their evaluation: Development of the Upper Level of the Behaviour Change Intervention Ontology', *Wellcome Open Research*, 5, p. 123. doi: 10.12688/wellcomeopenres.15902.1.

Michie, S., Thomas, J., *et al.* (2020) 'The Human Behaviour-Change Project: An artificial intelligence system to answer questions about changing behaviour', *Wellcome Open Research*, 5, p. 122. doi: 10.12688/wellcomeopenres.15900.1.

Michie, S., West, R. and Hastings, J. (2019) 'Creating ontological definitions for use in science', *Qeios*. Available at: https://doi.org/10.32388/ygif9b (Accessed: 17 November 2020).

Mikolov, T. *et al.* (2013) 'Efficient Estimation of Word Representations in Vector Space', *arXiv:1301.3781 [cs]*. Available at: http://arxiv.org/abs/1301.3781 (Accessed: 31 July 2021).

Neuhaus, F. (2018) 'What is an Ontology?', *arXiv:1810.09171 [cs]*. Available at: http://arxiv.org/abs/1810.09171 (Accessed: 25 August 2021).

Nunes, S., Sousa, R. T. and Pesquita, C. (2021) 'Predicting Gene-Disease Associations with Knowledge Graph Embeddings over Multiple Ontologies', *arXiv:2105.04944 [cs]*. Available at: http://arxiv.org/abs/2105.04944 (Accessed: 4 September 2021).

Obrst, L. *et al.* (2007) 'The Evaluation of Ontologies', in Baker, C. J. O. and Cheung, K.-H. (eds) *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*. Boston, MA: Springer US, pp. 139–158. doi: 10.1007/978-0-387-48438-9_8.

Pearl, J. (2018) 'Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution', *arXiv:1801.04016 [cs, stat]*. Available at: http://arxiv.org/abs/1801.04016 (Accessed: 5 September 2021).

Poveda-Villalón, M., Gómez-Pérez, A. and Suárez-Figueroa, M. C. (2014) 'OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation', *International Journal on Semantic Web and Information Systems (IJSWIS)*, 10(2), pp. 7–34.

Rector, A. *et al.* (2019) ‘On beyond Gruber: “Ontologies” in today’s biomedical information systems and the limits of OWL’, *Journal of Biomedical Informatics: X*, 2, p. 100002. doi: 10.1016/j.yjbinx.2019.100002.

Riegel, R. *et al.* (2020) ‘Logical Neural Networks’, *arXiv:2006.13155 [cs]*. Available at: http://arxiv.org/abs/2006.13155 (Accessed: 18 February 2021).

Salguero, A. G., Delgado, C. and Araque, F. (2009) ‘Easing the Definition of N–Ary Relations for Supporting Spatio–Temporal Models in OWL’, in Moreno-Díaz, R., Pichler, F., and Quesada-Arencibia, A. (eds) *Computer Aided Systems Theory - EUROCAST 2009*. Berlin, Heidelberg: Springer Berlin Heidelberg (Lecture Notes in Computer Science), pp. 271–278. doi: 10.1007/978-3-642-04772-5_36.

Serrano Nájera, G., Narganes Carlón, D. and Crowther, D. J. (2021) ‘TrendyGenes, a computational pipeline for the detection of literature trends in academia and drug discovery’, *Scientific Reports*, 11(1), p. 15747. doi: 10.1038/s41598-021-94897-9.

Smaili, F. Z., Gao, X. and Hoehndorf, R. (2019) ‘OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction’, *Bioinformatics*. Edited by J. Wren, 35(12), pp. 2133–2140. doi: 10.1093/bioinformatics/bty933.

Smith, B. *et al.* (2007) ‘The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.’, *Nature Biotechnology*, 25(11), pp. 1251–1255. doi: 10.1038/nbt1346.

Song, B. *et al.* (2021) ‘Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison’, *Briefings in Bioinformatics*, (bbab282). doi: 10.1093/bib/bbab282.

Thomas, P. D. *et al.* (2019) ‘Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems’, *Nature Genetics*, 51(10), pp. 1429–1433. doi: 10.1038/s41588-019-0500-1.

Vaswani, A. *et al.* (2017) ‘Attention Is All You Need’, *arXiv:1706.03762 [cs]*. Available at: http://arxiv.org/abs/1706.03762 (Accessed: 6 October 2020).

Völker, J., Fernández, S. and Sure-Vetter, Y. (2008) ‘Supporting the Construction of Spanish Legal Ontologies with Text2Onto’, in, pp. 105–112. doi: 10.1007/978-3-540-85569-9_7.