Semantic Web 0 (2021) 1–0
IOS Press

Knowledge graphs for common-sense scientific question answering

Guy Aglionby [a,*] and Simone Teufel [a]
[a] *Department of Computer Science and Technology, University of Cambridge, United Kingdom*
E-mail: guy.aglionby@cl.cam.ac.uk

Abstract.
Knowledge graphs (KGs) can be used to structure the information necessary for a model to successfully answer questions. In this paper we specifically investigate the storage of common sense information that expresses properties of abstract concepts. Prior work has examined ontology design for specific kinds of common sense, but the general case is under-explored. We identify weaknesses in the structure of ConceptNet, the predominant resource for general common sense, and propose a new modular ontology for common sense – MOntCS – to store this information. MOntCS is designed to be suitable for structuring explanations for questions by limiting the complexity of concepts permitted. We draw on linguistic theory to ensure consistency and clarity in the relation set. We use MOntCS to structure the facts provided with WorldTree, a scientific common sense question answering dataset which originally stores information in tables, and release this as a resource for knowledge graph-augmented question answering. We show that, with an existing knowledge graph reasoning model, using this knowledge graph gives higher accuracy compared with three competitor knowledge graphs. We carry out an ablation study to identify which relation types most impact question answering performance, and study which properties of knowledge graphs correlate with higher performance. We provide empirical evidence for claims made in prior work that taxonomic relations may not be useful for common sense reasoning.

Keywords: Common sense, Knowledge graph, Explanation, Question answering

1. Introduction

For a model to successfully answer questions it must have access to sufficient relevant information to make a decision. A common approach is to use latent representations of knowledge held in the parameters of large pre-trained masked language models (MLMs) [1, 2]. Another is to provide models with declarative knowledge, for example through inclusion of relevant text as input to the language model [3], or with the addition of a module to process graph-structured data [4, 5]. Several authors report that this approach improves model performance [4–7]. Another attractive aspect of this approach is that the use of such symbolic information held outside the MLM can potentially act as the basis of human-interpretable explanations for the system's decisions. These explanations can be used by developers to verify whether the model is combining facts in a way that entails the selected answer, and if not serve as a starting point for troubleshooting. In this paper we examine the behaviour of models provided with information extracted from different knowledge graphs for common sense question answering (QA).

Common sense QA, in contrast to factoid QA, is concerned with concepts and events as they *generally* hold in the world, instead of particular examples of concepts. For example, rather than asking about the size or population

*Corresponding author. E-mail: guy.aglionby@cl.cam.ac.uk.

ConceptNet	Proposed Framework
(knife, *used for*, slice) (knife, *capable of*, cut)	(slice, *instrument*, knife) (cut, *instrument*, knife) (slice, *agent*, person)
(bird's foot, part of, bird) (feather, part of, bird) (bird, has a, wings)	(bird, has subcomponent, foot) (bird, has subcomponent, feather) (bird, has subcomponent, wing)
(recycling, *capable of*, reducing waste reaching landfill)	(recycle, *cause*, reduce waste) (waste, *at location*, landfill)

Table 1
Comparison between facts expressed in ConceptNet, a commonly-used common sense knowledge graph, and in the proposed ontology.

of a particular city, common sense questions focus on properties usually held by cities. Common sense information is assumed to be known by most people as a result of their experience interacting with the world, and so is rarely stated explicitly [8]. Contemporary work on common sense ontology design has focused on specific domains of information; for example, ATOMIC [9] categorises different kinds of 'if-then' relationships between specific events. The design of ontologies for more general, open-ended common sense reasoning has been so far under-explored, and this is where our current paper's focus lies.

ConceptNet [10] is by far the most commonly used ontology for general purpose reasoning. It expresses relationships between natural language phrases that represent an aspect of the world. Relationships use one of a fixed set of relation types, which includes a generic *RelatedTo* type for use when no other one is appropriate. Each fact within the graph is expressed as a (subject, relation, object) triple.

We identify two aspects of ConceptNet's design that reduce its ability to structure common sense data meaningfully. First, there is systematic ambiguity in the relations that are available, meaning that more than one relation type can be used to express some kinds of relationship between concepts. This creates redundancy. For example, it expresses that knives are both *capable of* cutting, and *used for* slicing. Both of these relationships make sense when expressed as a sentence, however it is the job of an ontology to abstract above different expressions of the same relationship type. Although these two relations certainly are not synonymous – for example, only one would be relevant for discussing the capabilities of an artist – the large degree of overlap in relationship expressed between the two is undesirable. We would expect an ontology to unambiguously provide a single relevant relation this scenario. Second, overlap in meaning between relations may lead to inconsistency when annotating data and may make learning representations of relations more difficult. This was also noticed in prior work, and [7] were able to merge some relations with particularly high overlap while maintaining reasonable consistency within each cluster. What remains in this case are relations that do not overlap enough to merge, but that still overlap to an undesirable degree.

The second difficulty is that there is no structuring in the ontology beyond simple concepts and relations, and in particular no restriction on what a concept can be. Concepts are therefore expressed using many kinds of grammatical construction and at many different levels of specificity. One fact expresses that recycling has the capability of 'reducing waste reaching landfill', which we suggest is too specific as a concept. The core idea expressed here is that recycling reduces waste; the additional specification of where the waste is is not necessary, and should be expressed as a separate triple. It is difficult to ensure that a relation set is appropriate when used with concepts at many different levels of specificity We hypothesize that this is why the majority of relations in the graph to take a generic value. We expect that high quality explanations would contain clearly-defined facts about concepts that are defined at the same, or similar level of complexity as each other. As a result, we expect that the average explanation quality possible in ConceptNet is limited.

We propose a new ontology *MOntCS* – Modular Ontology for Common Sense – for this type of knowledge which addresses these issues. We express events as structured nodes[1], which combine multiple atomic concepts into one according to a set of semantic rules. The rules place an upper limit on the complexity of nodes. This approach is

[1]We refer to instantiated classes as nodes, following knowledge graph terminology.

similar to GLUCOSE's [11], where Mostafazadeh et al. allow events to be specified only via filling dependency slots for a verb. We provide clear definitions for remaining relations in MOntCS, which are derived from work in lexical semantics, semantic roles [12, 13], and narrative understanding [14]. We believe that these choices will reduce sparsity within relations, and that a specific and well-defined relation set is a valid way forward, as it retains the possibility for future integration. During construction we identified specific areas of difficulty when trying to represent this kind of knowledge; we will report here how we have resolved them.

Scientific question answering requires common sense understanding about general properties of many different concepts, and is a challenging task for models [15]. WorldTree [16] is a scientific QA dataset which, in addition to multiple-choice questions, provides 62 semi-structured tables containing the facts judged necessary to answer them. We express these facts in our ontology, using this process both to tweak the design and to verify its suitability for this kind of information.

To evaluate the ontology, we run a question answering model on WorldTree and compare performance when different knowledge graphs are used. We find that, with a model that most reflects the impact of the graph, MOntCS outperforms three alternative knowledge graphs. We also conduct an ablation experiment to discern the impact of each relation family.

Our contributions in this paper are as follows:

- We propose a new ontology MOntCS for structuring common sense information used in question answering. We explicitly design the ontology for suitability in expressing explanations for answers (§3).
- We present a translation of WorldTree's set of facts into a knowledge graph, and describe our annotation methodology (§4).
- We evaluate question answering performance on WorldTree, and find that the resulting graph is more useful for question answering than three alternatives. We also find a disproportionate reliance on language models in an existing QA system (§5).

2. Related work

2.1. Semantic resources

2.1.1. ConceptNet

ConceptNet [10] is perhaps the most frequently used knowledge graph for common sense reasoning applications. It aims to store the world knowledge required to understand language; although it has relations that facilitate this, much of the information actually stored in taxonomic or lexical semantic relations rather than common sense ones [9, 11].

The ontology of ConceptNet was originally developed in parallel with a data collection process: some relations directly relate to template-filling sentences; others derive from analysis of free-form input [17]. It supports five part-of-speech-based classes, although in practice most nodes are assigned to a sixth generic class.

In general, relations are free to hold between two nodes of any class. Although the range and domain of some relations are defined (for example, only noun-types have a *CapableOf* relation, that must be filled with a verb-type), they are not enforced in practice. Node names consist of free form text, and from version 5.5 on, terms are not lemmatised [10]. Instead, inflected terms are linked to each other using a specific relation type. In cases where more complex concepts are used, ConceptNet does not mandate that they are connected to the more general concepts they are related to. For example, the third row of table 1 contains a concept 'good for the environment'. It is not clear which relation should be used to link this with 'environment'; indeed no such triple exists in the graph.

2.1.2. WordNet

WordNet is a lexical database. At its most basic level, it represents word forms and collocations via membership of synsets [18]. These are collections of particular senses of lexical units with the same meaning, accompanied by a definition of the set. Synsets can also have conceptual links to others. Most relationships hold between synsets of the same part of speech, with the exception of "derivationally related forms". Nominal relationships include

part-whole (*meronymy*) and hierarchical "is-a" (*hypernymy/hyponymy*), while verbal relationships mainly concern manner (*troponymy*).

Since lexical units are stored in their base form, WordNet provides a tool *Morphy* with which to remove inflections and assist with search. Although some prior work has found WordNet to be helpful for question answering [19], its use is limited.

2.1.3. Aristo Tuple KB

Aristo Tuple KB[2] [20] is a collection of (**subject**, *relation*, **object**) triples relating to elementary school science. Triples are extracted using OpenIE from a collection of domain-relevant sentences. Unprocessed, these are too noisy for use, so they propose a multi-stage process using heuristics and crowd annotation to filter the dataset. Part of this process is to merge different surface forms that express the same concept, and to collapse verb phrases expressing the same relationship.

Naming relation types by lifting verb phrases from text raises the difficult question of how to handle polysemous verbs like 'have', which if used as a relation would encompass many different types of connection [20]. They propose keeping these relationships, and using the senses of the two joined concepts to aid with disambiguation.

The final graph has 1605 unique relations, where the top 15 are used in 50% of triples and 81% of which are used under 100 times. They find that it has a recall of 23% at 80% precision over a reference set of facts.

2.1.4. Event-centred datasets

ATOMIC [9] is a crowdsourced commonsense knowledge graph of *if-then* relationships between events. Crowd workers are prompted with an event, and are asked to provide events which are likely to precede or follow the prompt. Events are defined as verb phrases and are collected as free-form text, averaging less than five tokens each and in most cases (85%) appearing only once within the dataset.

GLUCOSE [11] concerns causality of events within short stories. Crowd workers are asked to relate events within a short story to each other in the context of different types of causal relationship. In some cases workers extend this to plausible but unstated events. The authors found that requiring antecedent and consequent events to be annotated separately, and a relation to be chosen from a fixed set, led to much simpler evaluation and processing compared with allowing free text. Event descriptions are further constrained via simple syntactic rules. Structuring explanations in this way also facilitates the construction of generalised rules not rooted in the given context, although how to complete this translation automatically is an open question.

2.2. WorldTree

WorldTree [16] is a question answering dataset built from elementary school-level multiple choice exams. The advantage of this domain is that selection of the correct answer requires the combination of multiple facts, while the language used is linguistically simple. This makes for a challenging reasoning task expressed in a way that can be comprehended by current models [15].

Adjective	Thing	Is/has	Value	Attribute
	floodplains	are	flat	in shape
	carbon dioxide	is	colorless	
	amphibian	has	smooth	skin
fresh	food	is	safe	to eat
rich	soil	has	a high	number of nutrients

Table 2
Except from the WorldTree 'properties-things' table.

WorldTree is released in conjunction with a *table store* – a collection of 62 manually-constructed tables containing the information required to anwer each question in the dataset. Each table represents a particular type of relationship

[2]We use version 5 (March 2017), courtesy of Allen Institute for AI.

or object property, and contains different columns as appropriate. The type of data stored within a column is intended to be consistent, although the variety of facts stored and the development of the schema alongside the annotation effort makes this difficult. For example, table 2 shows an excerpt from the 'properties-things' table in which two different relationships are expressed. The first two rows express adjectival properties of floodplains and carbon dioxide respectively, whereas the primary relationship expressed in the third row is that a meaningful constituent of amphibians is skin. Additionally, the entry types in the attribute column are not consistent.

2.3. Use of KG in QA

Knowledge graphs are frequently used with models as a source of information for natural language understanding tasks, including multiple choice question answering. ConceptNet is most commonly used as the base knowledge graph, a subset of which is chosen for computational reasons. Inverse relations are normally added, and in some cases relations are collapsed into each other to address the skew in distribution [7]. Early work incorporated embeddings of individual triples [19, 21]; recent work has extended this to representations of paths [4, 7] and subgraphs [5].

Simple techniques are usually used to identify relevant portions of the graph for each question and answer candidate pair. Linking entities mentioned in the question, or in an answer candidate, to entities within the graph is done using (lemmatised) lexical overlap [7, 19]. One-hop neighbours of these are retrieved and encoded in the simplest case. Path- and subgraph-based approaches extend this by finding numerous multi-hop paths between these linked entities, building up a set of intermediary nodes until a maximum number is reached or until path length exceeds a limit. Limiting path length is particularly important, as longer paths are more likely to relate concepts via spurious paths ('semantic drift') [22]. The extracted graph is referred to as the *schema graph*.

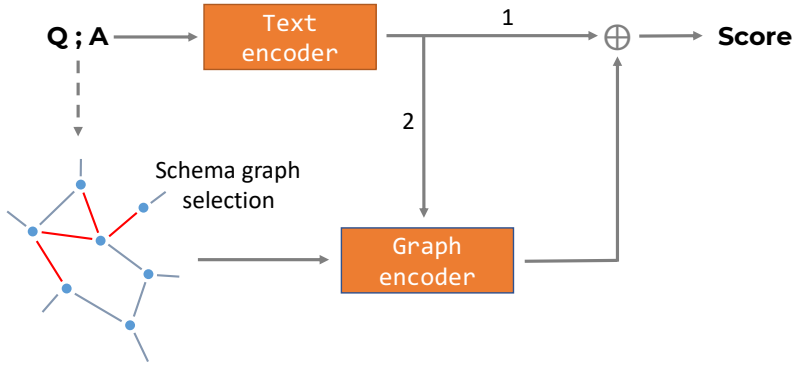

Fig. 1. Generic architecture of models which use text encoders and graph encoders. '1' shows the direct use of the text embedding in the final representation. '2' denotes some use of the text embedding in the graph encoder.

Figure 1 demonstrates the overall architecture of KG-augmented question answering systems. The schema graph is either encoded whole, for example using a GNN [5], or through a combination of embeddings of multiple paths within it [4, 7]. A language model is usually used to represent the semantics of the question in this process, for example to weight path embeddings in a weighted sum [4, 7] or as a pseudo-node in a GNN [5]. The resulting schema graph embedding is then concatenated with an embedding from the language model to give an overall score for the question and answer candidate pair.

2.4. Semantic roles

Semantic roles define the different kinds of relationships that verbs have with their arguments [12, 13]. In contrast to syntactic relationships such as "object" or "subject", they carry additional information about the event which is implict in the verb semantics. For instance, while agents are often expressed as syntactic subjects, they aren't always; in sentences in passive voice, agents are expressed as prepositional phrases headed by "by". Reversely, it is not always the case that a subject is an agent: in the sentence "he feared repercussions", the subject is a passive

experiencer of a state or an emotion, rather than an active participant in an event. Taking the sentence ‘we cooked food with the stove’, it is clear that the verb has different relationships with the two arguments “food" (realised as a direct object) and “stove" (realised as an indirect object). “Food" here is the patient, whereas “stove" is the `instrument` used in the cooking event. Semantic roles are useful in generalising above sentence structures: the semantic roles in the passive sentence ‘the stove was used by us to cook food’ remain the same, while the syntactic structure changes.

There are many approaches to defining a set of semantic roles, including those taken by FrameNet [23], PropBank [24], and VerbNet [25]. VerbNet uses a common set of roles across all of the groups of verbs classes it recognises, whereas PropBank and FrameNet don’t operate with a common set of rules. PropBank defines defines different types of role per verb[3], and Framenet different types of roles per groups of verbs (called a "frame" in FrameNet), although in both lexica there is some consistency between roles in some cases. No formal general description of a role is given, but the lexica rely on human’s intuitions when interpreting them. All definitions are given only in the context of a frame. From this sample of approaches it can be seen that there is little consensus about the appropriate granularity of roles, so we had to choose our own for our knowledge graph, from the vocabulary available in the literature.

Using PropBank-style roles would give a large set of sparsely-populated relations, which is undesirable both from a design perspective and a modelling one as it is difficult to learn representations from few instances. We therefore used a general set of relations similar to VerbNet’s (although there is also disagreement about suitable role sets at this level of granularity).

3. Ontology design

Our proposed ontology is designed with the goal of structuring the information required to answer common sense questions. We envisage models using the ontology both as input, to provide relevant information to aid in answering questions, and as a medium to structure explanations. We take the types of facts in WorldTree to be indicative of what is required in the general case, though are conscious not to introduce designs which are overly specific to the science domain. Knowledge graphs expressed in this ontology are expected to be used with systems that can separately interpret the context (in our case, a question), select relevant nodes from the graph, and manipulate these to complete a task.

We define relations to hold between nodes when it is plausible that the type of relationship exists, rather than if it strictly holds. Common sense facts by nature have many exceptions: not all dogs can bark, and not all bears are brown. These particular facts however are true frequently enough to be useful when answering common sense questions.

The impact of this choice is three-fold. Firstly, it simplifies the ontology, removing the need for either more fine-grained relations or meta-relations to express more complicated constraints on a relationship. Secondly, it simplifies annotation, by not requiring an annotator to specify, for example, precisely which subtypes of bears are brown. Thirdly, it requires a system using this knowledge graph to be capable of defeasible reasoning, making use of the context in which it is deployed.

We design the ontology to promote the construction of dense knowledge graphs. This is important when the graph is used with KG-augmented systems, at most are only able to operate on a subgraph of the overall graph. As a graph grows denser, it becomes easier to select relevant data that may otherwise require many hops to reach from the starting nodes. Using an increasing number of hops to select a subgraph quickly results in a large amount of noise [22].

We impose two constraints to ensure that facts expressed using the ontology are at similar levels of specificity. The first is that we do not include a default relation type to join nodes where no others apply. This is in contrast to ConceptNet, which uses a general *RelatedTo* relation in 78% of cases. Instead, we express complex relationships as multiple, simpler facts.

[3]For example, PropBank defines the roles ‘consumer/eater’ and ‘meal’ for the verb ’eat’.

The second constraint is that we require that node names follow a simple syntactic grammar. The node's class is then based on the parts of speech present in the node name. This ensures that node names do not exceed a certain level of complexity, and aims to improve the generality of the resulting knowledge graph.

3.1. Classes

Similarly to ConceptNet, our base set of classes are based on parts of speech (verb, noun, and adjective), but also include qualifiers. These are distinguished mainly for ensuring consistency during annotation. Nodes that take these classes have their names lemmatised.

We also include classes for each type of structured node, which are constructions to allow the controlled expression of more complex concepts. The class of a structured node is defined by the POS tags of the words within it.

As all classes are based on (combinations of) part of speech tags, the hierarchy is flat. Taxonomic relationships are therefore expressed using a specific relation type rather than through class hierarchy (§3.3.2).

3.2. Structured nodes

Some common sense facts cannot be expressed using bare nouns, adjectives, or verbs; instead, they require that these base classes are combined in some way to express more complex concepts. For example, to express that the kind of change that acid can cause is a *chemical change*, we must be able to construct a node for this which is distinct from a generic *change*.

Structured nodes allow these concepts to be expressed while placing a limit on possible complexity. Not controlling complexity at all would result in inconsistently structured nodes, which may be difficult to fit with the fixed relation set and would be confusing when included in an explanation. Even if the structure is controlled, allowing the expression of too-specific nodes will reduce the generality of the graph. We use a grammar to limit nodes to simple phrases, which broadly reflect verb phrases and noun phrases. This is similar to the approach taken in GLUCOSE [11], where imposing syntactic constraints increased inter-annotator agreement and aided evaluation.

We define noun phrases to consist of an optional quantifier, zero or more adjectives, followed by one or more nouns. Verb phrases require a verb and one or both of an agent or patient, which themselves must be noun phrases. We create a class for each of the possible permutations under these conditions.

3.2.1. Structural relations

To increase connectivity, and therefore density, within the graph, we require that structured nodes have relations which join them to each of their constituent words. This family of seven relations is similar to the structural links proposed by Woods [26, p.35]. A straightforward benefit of this requirement is reducing the likelihood that these nodes become disconnected from the rest of the graph. A larger benefit is their impact on the schema graph selection process, the first step of which is to select nodes which are relevant to the question or answer. If this does not manage to pick out some particularly relevant structured node, the subsequent graph exploration process may not find it even if one of its constituent words was initially identified. Even if a relevant structured node is identified, accessing more general information about the concepts mentioned may be difficult for the same reason. Structured relations therefore reduce reliance on the first step of the schema graph selection process to identify good nodes, and makes it easier to transition between specific and general facts.

The structured relations available, their domains, and their ranges mirror the classes in the ontology. The `structural noun`, `structural noun compound`, `structural quantifier`, and `structural adjective` relations are used with noun phrases. The domain of these relations relates to the specific type of noun phrase represented by different classes – for example, a compound noun class should have `structural noun compound` relations. Classes involving adjectives should use both `structural adjective` and `structural noun` instead. Each relation's range is restricted to basic part-of-speech classes; in particular both structural noun relations cannot be filled by another noun phrase.

Verb phrases use the `structural verb`, `structural agent`, and `structural patient` relations. The domains and ranges of these relations are defined in the same way as for noun phrase-related relations; in particular the range of `structural verb` is limited to verbs, while the other two may be nouns or noun phrases.

3.3. Relations

Our ontology contains four additioanl families of relations that express common sense relationships between nodes.

3.3.1. Verbal relations

A key contribution we make is to apply ideas from the study of semantic roles (§2.4) to knowledge graphs, with the aim of minimising ambiguity between types of verbal relation. The seven relations we use are listed in table 3.

Relation	Definition
Agent	The volitional or non-volitional causer of an event.
Patient	The undergoer, experiencer, or entity moved by the event.
Instrument	An entity used in an event.
Result	The end product of an event.
Beneficiary	The entity that benefits from the event.
Source	The origin of an entity in a transfer event.
Goal	The destination of an entity in a transfer event.

Table 3
Verbal relations used in the ontology, with definitions adapted from [27].

Understanding the roles of entities used in or produced by an event is crucial for understanding what happened, and knowing the impact of an event on an entity is useful in reasoning about what may happen next. All relations express high-level information past simply listing typical values for each grammatical dependency that a given verb could have. In particular, the 'instrument', 'result', and 'beneficiary' relations express information that is likely to be assumed as shared knowledge and not explicitly stated [8], which is one way to define common sense knowledge.

For simplicity, and to ensure that there are no relations that are disproportionately (in)frequent, we have also removed some distinctions that are usually made in semantic role sets. We do not distinguish between an 'patient', which directly experiences a change of state in an event, and a 'theme', which is similarly necessary for the event but is not changed by the event. We also erase the distinction between entities which intentionally and non-intentionally cause cause an event, and those which are primary or secondary causers.

Our relationships have the advantage over those in ConceptNet in that they are more principled and less redundant, although it is of course possible to express properties and relationships similar to the ones we chose, within ConceptNet's relation vocabulary. For example, ConceptNet expresses that a knife is both `capable of` cutting and `used for` slicing. As the two verbs express essentially the same event the same relation should be used, which in our ontology is 'instrument'. This additionally asserts the useful information that knives tend not to be agents, i.e., they do not act in events autonomously without another entity present which is the cause of the event.

3.3.2. Taxonomic relations

Taxonomic relations express lexical relations holding between concepts, such as synonymy and hyponymy. Lexical relations specifically hold just between words and other words, and do not relate to the concepts that they denote. For example, WordNet [28] defines one sense of 'hypothesis' to be synonymous with 'theory', and these to be hyponyms, or kinds of, 'concept'.

It is not clear that basic relations such as these could be useful for common sense reasoning, as has been highlighted by Mostafazadeh et al. [11]. However, we include four commonly known taxonomic relationships in case the additional connections between concepts are useful when extracting information from the graph. We summarise these relations in table 4.

The closest we come to in our ontology to a generic relationship type is 'similar to'. It is necessary to include this as two things may be similar enough to be notable, but too dissimilar to include in the set of synonymous things. For example, mass and weight are related concepts, but particularly in a scientific context it would be incorrect to say that they were the same thing.

Relation	Definition
Synonymy	Two concepts mean the same thing, or are extremely similar.
Similar to	Two concepts are very similar, though not similar enough to be synonymous.
Antonymy	Two concepts are opposites.
Kind of	Hyponymy; one concept is a kind of another.
Instance of	Where a concrete entity is an example of a general concept.

Table 4
Taxonomic relations used in the ontology.

When it comes to hyponymy relationships, our definition is permissive and even allows connections across part-of-speech boundary ('far' is a kind of 'distance'). We also do not distinguish between hyponymy and quasi-hyponymy [29, 30], as this distinction does not in our opinion help in common sense scientific QA, where even rough connections between concepts are often sufficient to guide the reasoner towards the correct answer.

The `instance of` relation is different from `kind of` in that its domain concerns instances, i.e., concrete entities. It is used therefore to distinguish the relationship 'city' has with 'Cambridge' and 'settlement' – cities are kinds of settlement, whereas a particular example of a city is Cambridge.

3.3.3. Affective relations

We use two relations to express positive or negative impact of events and concepts on others. These relations are useful for two reasons. Firstly, they can be used to represent specific relationships between objects in an abstract way: for example, the different high-level impacts of recycling, pollution, and soil erosion on the environment. An alternative would either require specific relations to be introduced, or multiple triples to be created that express each concept in detail. These approaches may still fail to encode that one is a positive concept while the others are negative. ConceptNet expresses recycling's impact via a 'has property' relation of 'good for the environment', which we argue is too complex as a concept. ConceptNet also contains `desires` and `not desires` relations, however their domain is explicitly limited to conscious entities [31].

Secondly, affect is central to understanding narrative stories, which can be modelled as a sequence of positive, negative, and neutral states of protagonists [14]. Some scientific common sense questions concern events, which are simply a point in time within a narrative and so can be modelled in a similar way. These relations therefore help ensure that the ontology is extensible to other domains.

3.3.4. Other relations

We include six additional relations outside of the four main families, which share similar definitions to their equivalents in ConceptNet. These are listed in table 5.

Relation	Definition
Causes	Possibility that an event or entity causes, or helps in causing an event or phenomenon.
Has property	An attribute (usually adjectival) of a concept (usually nominal).
At location	A location where either an event can happen or an entity can be found.
Has subcomponent	A meronymic relationship; one concept has another as a constituent, part, or subsection.
Linked by morphology	Two concepts are linked via derivational morphology, for example verb nominalisations.

Table 5
Relation types outside the four main families.

Our definition of the causal relationship collapses causes, causal factors, and possible causes into a single relation type. Meronymy is an interesting and complex taxonomic relationship with several alternative interpretations in the literature [30]. We leave a detailed examination of which types could be used in expressing scientific facts to future work.

3.4. Grammatical phenomena covered

We attempt to cover the variation of natural language phenomena to a certain degree, but it is well known in the literature that there are many limitations to using a simplistic triple representation, which is theoretically insufficient to represent complex natural language phenomena [32]. In some cases, our methods for handling these phenomena are informed by the annotation of WorldTree discussed in §4.

Arguments All arguments for all verbs, including intransitive, transitive, and ditransitive, are connected using verbal relations (§3.3.1). Intransitive verbs are simple to express: a node exists for the verbal concept, and triples can be created with the appropriate verbal relation for each subject encountered in the tables. Structured nodes can be constructed in a similar way.

For transitive, ditransitive and other more complex verbs involving more than one argument, we choose a modular expression that uses multiple, independent, and simpler triples, while also making use of structured nodes. We observed that few complex facts expressed with complex verbs lost all their meaning when modularised in this way.

In the transitive case, take the example of 'birds eat seeds'. It is useful to express this as generally as possible, and there are two ways of doing this. The first is to create a structured node of the verb and agent, and create a triple ('birds eat', patient, 'seed'). The second would be to instead create a structured node involving the patient: ('eat seed', agent, 'bird'). Generally we expect that the most generic of these possibilities be used to express the relationship, although in this case as birds eat many things, and many things eat seeds, both are equally valid. A secondary consideration when choosing between possibilities is whether one is likely to relate more closely to the kinds of questions that are asked. These judgements are made at construction time by an annotator. Only if it was necessary to express the entire fact within a single concept should a structured node involving both the agent and patient be created. For example, we could express that birds typically eat seeds in a bird feeder by the triple ('bird eat seed', at location, 'bird feeder').

We demonstrate how to express ditransitive verbs with three necessary arguments with the example 'bakers give bread to customers'. A structured node containing the agent, verb, and patient should be created, and this underspecified node used in a triple that expresses the complete fact. This gives a final triple ('bakers give bread', beneficiary, 'customer'). Relationships that involve the fully specified ditransitive verb, for example 'green plants provide food for animals by performing photosynthesis', cannot be specified in our ontology. However, we capture most of the meaning by specifying that that an agent of 'photosynthesize' is 'green plant', 'photosynthesize' results in 'food', and that the structured node 'animal eat' has a patient 'plant'.

Modification Noun modification includes but isn't limited to, prepositional phrase modification, adjectival modification, and noun compounding. Structured nodes explicitly allow for the last two kinds of modification.

We do not explicitly handle prepositional modifiers due to the substantial added complexity doing so would bring to structured nodes. Taking the example 'particles in objects are a kind of matter', if we were to create a node 'particles in objects' we would require structural relations in the same way as with verbal structured nodes: one relation each to indicate the head and object of the prepositional phrase. It would also be necessary to indicate the preposition used, which if accomplished in the same way as with verbs would create nodes for each preposition handled. Unlike verbal concepts, which have meaning even in the absence of arguments, standalone prepositional concepts would be meaningless.

We find that many prepositional modifiers can be expressed in a different way. In some cases the specification is not important: the fact 'particles in objects are a kind of matter' is fully expressed by two independent triples ('object', has subcomponent, 'particle') and ('particle', kind of, 'matter'). In other cases, we collapse prepositional modifiers into a compound that, although maybe not common or even obviously interpretable, still likely carries enough meaning to be useful for a task. For example, we create a noun compound 'ecosystem role' to represent 'role in the ecosystem'.

Where this is not possible, we attempt to express the specification as its own triple. For example, to express the locative preposition in 'a coal mine is a source of coal under the ground', we specify ('coal mine', at location, 'underground') and ('coal mine', has subcomponent, 'coal'). The disadvantage of this approach is that the second triple taken alone does not specify the location of the coal mine, although in practice this is unlikely to be important.

Additionally, we are unable to express the fact that 'younger rock is usually located under older rock', instead storing that they are located in the same place.

In our annotation we are able to closely approximate most relationships using the methods described. While relaxing a specification made in a factual statement is likely to make the assertion incorrect, doing so for common sense knowledge may only render the fact under-specified. Additionally, we find that not providing an explicit mechanism for prepositional modification (and arguments) forces annotators to create facts that are more general than if such modification could be expressed.

Negation The classic problem in any natural language ontology is of course how to express negation. Although we do not find many negations in the WorldTree data, this is a theoretical question that deserves discussion and treatment. Ideally, we would like an open world interpretation of our ontology: statements not explicitly mentioned can either be true or false, until more information comes in. Abstaining from making a decision is important because it is infeasible to include all possible information in a knowledge graph, meaning that making a strong decision on the basis of missing information is unwise.

To express negative information within the syntax of our ontology in an open world scenario, we could create a separate, negated version of each relation to be used when something *not* being the case was explicitly stated. However, this would represent a drastic increase in the number of relations in the graph for relatively little gain, given the low number of statements of this type. The sparse population of these relations may also make learning representations of them difficult.

Given that we can only add positive information to the ontology, our approach is to search for an appropriate antonym given a negative fact, and express both the positive fact and the antonymy relationship. In cases where this is not possible, we aim towards a situation where the lack of the link becomes conspicuous by its absence. For instance, expressing that the Moon does not have an atmosphere can be done by instead expressing that the Earth, Venus, Mars and Jupiter do. This is an imperfect solution that tries to simulate an open world situation but still closely resembles a closed world assumption.

Subjunction Relationships between clauses expressed by subjunction is too complex to express in our ontology. This includes if-then constructions. For example, within 'if a leaf falls off of a tree then that leaf is dead' the main clause can be expressed as the triple ('dead leaf falls', at location, 'tree'). Here, the condition is implied by only recording a triple that represents the world when it is true. The ways that such relationships are expressed depends on context and concepts involved. We were therefore forced to manually translate all such examples into suitable triples (see §4.2).

Quantification Quantification in our ontology concerns modifications to nouns to express their quantity, including expressions like 'few organisms', and 'less bacteria'. This modification is distinct from logical quantification – we do not discuss properties that hold for particular instances of a given concept, nor all instances of it. Some facts, like 'cold environments contain few organisms', require quantification to be fully expressed: simply constructing a triple ('cold environment', has subcomponent, 'organism') omits the primary aspect of the fact. As a result, we permit structured nodes to contain quantifiers for nouns. In this case a node 'few organisms' allows the full fact to be represented.

Another example where quantification is useful is in representing 'pasteurization reduces the amount of bacteria in milk'. The main clause can be neatly expressed as ('pasteurize', cause, 'less bacteria') and the prepositional phrase as ('pasteurize', patient, 'milk'). As with other structured nodes, we require that they have a relation with each of their components, including with the quantifiers themselves. Creating a node for all possible quantifiers is undesirable, as each has little intrinsic meaning. As such, we use a limited set of quantifiers, accepting in some cases that this results in ungrammatical nodes. For example, we express that hurricanes have large amounts of rain by the triple ('hurricane', has subcomponent, 'many rain').

Nominalisation In some cases, nominalisations occur in sentences, and our first step when dealing with such cases, considering the generalisability we seek, is to attempt to reformulate the clause with the corresponding verb. Whenever this is possible, we then simply revert to the rules for verbs, which is advantageous because our representations of verbs and their arguments is both relatively systematic and expressive (§3.3.1). Only where this is not possible do we include a link between a third concept and the nominalisation. For example, the fact that one gathers observations

from an experiment is not appropriately captured by ('observe', patient, 'experiment'), as observations have a particular meaning in this context. In all cases where a verb and its nominalisation, plus any other derivationally-related terms (e.g. 'snow', 'snowing', 'snowy'), exist within the graph, all are linked with a morphological-relatedness relation.

Comparatives Some facts represent comparisons between concepts, for example that a mountain is taller than a hill. These are impossible to model directly, as this would require a relation to be created expressing the comparative form of each relevant adjective. Instead, we create a triple for each concept in the comparison with the `has-property` relation, and use different quantifier of the adjective to express the relationship. In this case, we create two triples ('hill', has property, 'tall'), and ('mountain', has property, 'very tall').

4. WorldTree annotation

We annotated the facts in WorldTree according to our ontology to test whether it was suitable for expressing common sense and scientific information. Each fact was mapped to one or more (subject, relation, object) triples.

Annotation was completed in a semi-automatic manner. Some tables had columns that mapped directly into a triple representation and so could be translated automatically (§4.1); others did not and so required manual annotation (§4.2). In between these two extremes, some tables contained columns that could not be automatically handled but that were not populated in all rows. In some cases this did not preclude the row being partially automatically translated, allowing the additional information to be specified by hand later. In other cases automatic translation failed completely. Table 6 contains statistics about the translation methods employed across the dataset.

Translation method	Tables	Rows
Automatic	19	42.85%
Partially-automatic	13	1.21%
Manual	30	55.94%

Table 6

Breakdown of methods used to translate WorldTree table store. A table is considered completely automatically or manually translated if more than 90% of its rows are processed in that way. Percentages are given for rows independently of their table.

We post-processed the knowledge graph to ensure consistency and to fill in any relations missed during manual annotation (§4.3). This included adding missing structural links where they were missing – for example, if 'cut down tree' only had a nominal structural relation, a verbal structural property to the phrasal verb 'cut down' (not 'cut') should be added. We also selected just the largest connected component of the graph, as we expect that useful inputs for question answering models consist of connected graphs. This also reflects our requirement that good quality explanations should explain how each fact is relevant, which is not possible with a disconnected explanation graph. The final graph contains 5347 nodes and 13,965 triples, representing 4722 WorldTree facts.

4.1. Automatic translation

31% of tables were completely automatically translated, while 21% had a combination of automatic and manual translation. Fully-automatable tables encoded simple relationships, such as listing location-country relationships (which hemisphere a country is in), simple meronymic information (e.g. sub-processes within a process), and binary properties such as whether materials were magnetic. In the most straightforward tables, salient columns were identified and the relationship between them manually specified. Other tables included optional modifiers in other columns which were used to create structured nodes, and in some cases the relation used was chosen based on the value in a particular column. In many cases multiple valid terms were given within a cell; we created triples for all possible combinations of these within a row. Automatically applying rules in this way saves annotator time, increases reproducibility, and reduces the chance of annotation errors.

The 13 partially-automated tables expressed relationships like synonymy, hyponymy, and more complex meronymy. 36% of rows in these tables were manually annotated, as they contained terms which could not be directly used as node names in the ontology. The rest were automatically translated, although in 4.5% these some manual triples were also specified. Additionally, some rows were annotated with part of speech information to aid in automatic structured node creation, mainly to distinguish noun-noun and adjective-noun compounds.

4.2. Manual translation

The remaining rows were manually translated into a mean of 1.9 triples each (s.d. 1.2). These rows were in tables expressing complex relationships including those around events, affect, and if-then relationships; the final table containing 28 columns. The complexity of these relationships meant that each column had slightly different semantics within each row, precluding automatic translation.

To speed up annotation, we developed a meta-annotation method for structured nodes that allowed annotators to mark the part-of-speech tags of constituents. The tagged cell was processed with the same grammar used to limit node complexity (see §3.1) and structural relations automatically generated. Additional relationships implied by the node name were also generated – in figure 2, a valid agent for 'grow' and a property for 'thing' are also added. This method therefore ensures that all relevant relations are extracted from the annotation.

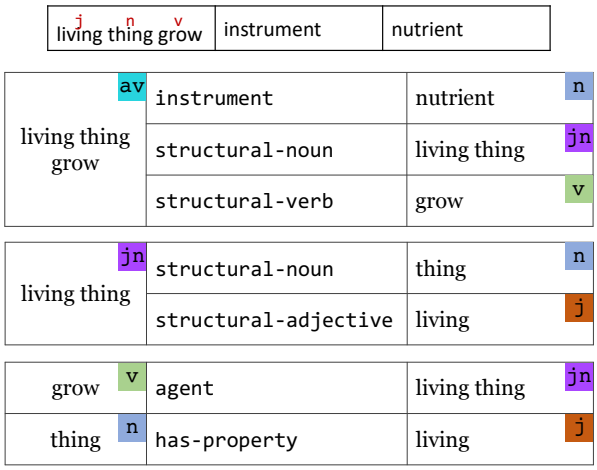

Fig. 2. By including POS tag information in a manually-annotated triple, structural and implied relations can be automatically generated.

4.3. Post-processing

After translation, we post-processed the nodes to check for annotation errors and to confirm that they conformed to the ontology. This included:

- Highlighting rows with non-stopword lemmas that do not appear in any nodes derived therefrom;
- Ensuring that all nodes have a class, and checking if the same surface form was used in more than one class;
- Verifying relation domain and range facets, including ensuring that all structured nodes have the required relations;
- Adding links between nodes with surface forms related by derivational morphology, for example between a verb and its nominalisation. We used Morphy for this [18, 33];
- Spell-checking.

We attempted to automatically label nodes that were not manually tagged with part of speech. In cases where either WordNet or other annotations unambiguously gave the part of speech, the corresponding class was used. As all verb phrases were manually annotated, all untagged multi-word nodes were noun phrases. If WordNet unam-

biguously identified both words in a two word node as nouns, or the first as an adjective and the second a noun, that classification was used. In other cases we used spaCy [34] to tag the node. All remaining nodes without a class were exported for manual classification.

We also highlighted cases where two nodes existed with the same name but different classes. In some cases, this was a conflict that needed to be resolved (e.g. whether 'living entity' is an adjective-noun pair or a compound noun, which we resolved as the former). In others this was genuine ('open container' is both a possible verb and noun phrase).

4.4. Graph statistics

For the final knowledge graph we remove classes and collapse nodes with the same surface form, and use only the largest connected component. This contains 5347 nodes (98.9% of total), with all discarded triples only using relations for synonymy or morphological relatedness. In the largest component, 55% of nodes represent nouns or noun compounds, 17% structured nodes involving a verb, 12% adjective-noun compounds, and 11% verbs (all include double counting where nodes have been collapsed). The distribution of relation types by family is given in table 7.

Relation family	**Count**
Structural	4959
Taxonomic	2599
Verbal	2578
Affective	311
Other	3468
Total	13,915

Table 7
Frequency of different relation families across all nodes.

We compare graph statistics of the graph against others in table 8. The *WT-automatic* graph is automatically constructed from the tables in WorldTree by creating an node for each row and each cell. A triple is created for each cell to connect it to its row, with the relation type defined by the cell's column title [35, §4]. The resulting graph is bipartite, so there is no triadic closure and the clustering coefficient is 0. We use the largest connected component for all graphs.

Graph	**# Nodes**	**# Edges**	**Clustering coef.**	**Density**	**Diameter**
ConceptNet	771,205	2,471,913	0.109	0.08	21
Tuple KB	44,912	282,550	0.112	2.80	9
WT-automatic	8880	16,484	0.000	4.18	20
MOntCS	5347	13,915	0.236	9.74	12
MOntCS / affective	5337	13,602	0.239	9.55	12
MOntCS / verbal	5243	11,311	0.153	8.23	13
MOntCS / taxonomic	4858	11,192	0.271	9.49	16
MOntCS / structural	4500	8668	0.076	8.56	16
MOntCS / other	4771	10,366	0.164	9.08	18

Table 8
Properties of different common sense knowledge graphs, including five data ablations of our graph corresponding to relation families.

ConceptNet is by far the largest graph, although it is also the least well-connected, with a low density[4] and clustering coefficient. This contrasts with MOntCS, which has a relatively low diameter and is at least twice as dense

as other graphs, allowing easier retrieval of useful concepts from those identified in a question. The automatically created graph is bipartite – 'row' nodes are only connected to 'cell' ones – and the resulting large diameter adds difficulty to the retrieval of relevant concepts.

The role of each relation family in structuring MOntCS can be discerned from the graphs where they are removed.[5] The decrease in clustering coefficient and density when removing both verbal and structural relations suggests their importance for traversing the graph. The increase in diameter and decrease in number of connected nodes in the structural case further reinforces this.

Although the removal of taxonomic relations also decreases the number of connected nodes, clustering coefficient actually increases. This suggests that other relations are sufficient to ensure connectivity between most remaining nodes, although because the diameter increases it is likely that some become more isolated. The increase in clustering coefficient also suggests that the removed vertices were not well-connected, and indeed 64% of the nodes lost were 'leaves' that only featured in one triple.

5. Evaluation

5.1. Methodology

To test our design hypotheses, we compare model performance on WorldTree when using our graph and existing alternatives. We additionally experiment with ablations of our graph, removing each family of relations in turn and analysing the impact on performance.

We select QA-GNN [5] as the model owing to its high performance on scientific question answering. It selects an answer by creating a representation for each choice and calculating a probability distribution over them. Each representation is composed from a masked language model (MLM) embedding of the question and candidate answer, and an embedding of the knowledge graph from a graph neural network (GNN) (see §2.3 and figure 1).

We report results using the original version of QA-GNN, as well as with two amended versions designed to isolate the impact of the knowledge graph on performance. The design of QA-GNN does not ensure that the graph is used in a meaningful way due to the MLM embedding being used directly in the representation used to score each example (see figure 1, label '1'). Language models are known to perform well on question answering tasks [1], and it is difficult to know the extent to which they drive performance in this model. As such, in the first ablation we do not concatenate this embedding to the final representation, and so the final score is only calculated from the GNN representation. However, because QA-GNN also includes this embedding within the GNN (figure 1, label '2'), the langauge model can still be trained. Our second ablation is therefore to freeze the MLM weights. In this scenario the only trainable weights are in the GNN, maximising the influence that changing the input graph has on performance.

We follow prior work [7] and construct schema graphs per seed from nodes found on paths between all pairs of concepts found in a question and in each of its answer candidates. As scientific question answering requires the combination of multiple facts which are often distantly related [15, 16], a path length of 2 as used in prior work is insufficient. Ensuring fairness in this scenario is difficult. Collecting nodes from all paths up to a maximum length would be noisier in larger graphs than smaller, from which small and focused subgraphs would be extracted. We are also careful not to bias towards smaller graphs by limiting the number of nodes extracted too extensively. We compromise by building schema graphs of 50 nodes each, corresponding to an average path length of 4.4 for ConceptNet, 4.2 for TupleKB, 7.2 for WT-automatic, and 6.2 for MOntCS. As a result the schema graphs extracted from the final two knowledge graphs are likely noisier than the others due to semantic drift [22].

TupleKB contains 1605 relation types, although their frequency is heavily skewed towards the more common ones. For computational reasons, we use the most frequent 200 and collapse all others in to a generic type. The WT-automatic graph has 171 relation types corresponding to the columns within WorldTree's table store. We follow prior work in collapsing some ConceptNet relations [7] to yield 17.

[4] All density values are given at e−4.

[5] The number of edges on ablations are not consistent with table 7 as other triples may be removed when a node is disconnected from the main component.

We use the same model configuration and hyperparameters in all scenarios. WorldTree has only 987 training instances, so we pre-train on the Open Book Question Answering (OBQA) dataset of scientific questions [36] before fine-tuning. The model is optimized with RAdam [37], with batch size 64 and L2 weight decay 0.01. Training is stopped after 10 epochs of no improvement on the development set. Each experiment is run four times with different random seeds and the results averaged. The language model used is a pre-trained `Roberta-large` instance [38] implemented with HuggingFace Transformers [39]. It is frozen for the first five epochs of training and then trained with learning rate 1e−3. The four-layer, dimension 100 GNN is trained at learning rate 1e−5. We find that applying GroupNorm [40] after each layer improves performance; we do not use batch normalization as we use gradient accumulation. Initial node embeddings for the GNN are derived from `Roberta-large` following the method in [4].

5.2. Question answering results

Graph	QA-GNN	− MLM. embedding	− Train MLM.
MOntCS	**62.69**	**56.49**	**46.93**
WT-automatic	55.77	49.14	35.14
TupleKB	60.58	54.04	21.20
ConceptNet	61.73	54.67	29.90

Table 9

Accuracy on WorldTree using QA-GN with four different graphs as input. Columns represent two successive model ablations on QA-GNN that emphasize the role of the graph in prediction.

Table 9 shows that MOntCS gives the highest performance in all three model settings, although this is unsurprising given that it is tailored towards the questions present in WorldTree. The standard QA-GNN model and the model with the MLM embedding ablation each give similar results across graphs, with the exception of the automatically translated graph. For all graphs, as successive model ablations are applied, performance decreases.

As only the GNN is being trained in the final column, this is the scenario in which changes to the input graph have the most impact. We note that the model was not developed with this in mind and so absolute performance is relatively low. However, what is more interesting here is the drop in performance for each graph compared with the original model. Not only does MOntCS perform the best, but it drops just 25% relative performance, compared with 37% for WT-automatic, 65% for TupleKB, and 52% for ConceptNet. This is the strongest evidence of the suitability of the proposed ontology in itself, over existing external graphs, and over automatic translation methods.

TupleKB performs below random chance (25%) in the final column, despite being a science-focused knowledge graph. This is likely due to its provenance – it is built from OpenIE extractions from text, which limits the type and complexity of relationships gathered.

5.3. Graph ablation study

We perform an data ablation study to evaluate the impact of different relation families on performance. We remove relations in each family from the graph in turn, re-extract schema graphs, and re-train models in the same way as previously. Results are shown in table 10.

We focus on the final column, where a change in graph contents most impacts model performance. Here, all but one data ablation reduces performance. The fact that performance increases when taxonomic relations are removed from the graph suggests that they encode information which the model does not need, and so could be classed as noise in the schema graph. Taxonomic knowledge may instead be provided by the MLM-derived representation used in the GNN [41]. An alternative explanation is that taxonomic triples not useful for retrieving relevant information from the knowledge graph. Table 8 shows that with these removed, clustering coefficient actually increases, and there is only a small decrease in graph density. We conclude therefore that taxonomic links are not necessary to ensuring high connectivity between nodes. This result provides empirical evidence for the suggestion that explicit

Graph	QA-GNN	− MLM. embedding	− Train MLM.
MOntCS	**62.69**	56.49	46.93
MOntCS / affective	62.12	60.19	44.66
MOntCS / verbal	59.81	**60.58**	39.14
MOntCS / taxonomic	62.45	55.92	**47.26**
MOntCS / structural	**62.69**	57.84	39.81
MOntCS / other	60.29	55.24	36.88

Table 10
Accuracy on WorldTree using QA-GNN with MOntCS, plus five separate data ablations. In each ablation, relations are removed from the graph and the model re-trained.

taxonomic knowledge is of limited use for common sense reasoning [11], despite it being a focus of prior data collection [42].

The four other ablations give reduced performance in the final column, confirming their imporatance in the ontology. The importance of affective relations is highlighted by the disproportionate 2.27% drop in performance when they are removed, as they make up just 2.25% of edges and 99.81% of nodes remain after their removal. In table 10, in the final column there is a strong but not significant linear correlation (Pearson's $r = 0.653$) between performance and the number of edges of each graph, and nearly significant correlation for density ($r = 0.763$, $p < 0.08$) and clustering coefficient ($r = 0.801$, $p < 0.06$). We conclude that highly inter-connected graphs are likely to aid performance in this modelling setup.

Extending the analysis to all nine graphs supports this conclusion. Although the structure of the graphs are not directly comparable, performance significantly correlates with density ($r = 0.833$, $p < 0.006$) and graph clustering coefficient ($r = 0.675$, $p < 0.05$).

6. Discussion

Models are increasingly evaluated not just on performance, but on their ability to provide explanations for the choices they make. We design MOntCS to be a suitable medium for expressing explanations in the common sense question answering domain. This suitability comes from relations with appropriately specific definitions, which ensure clarity in the meaning of each triple. Clear meaning of triples facilitates comparison between them, which is important in explanation evaluation when considering the inclusion (or not) of a triple in an explanation. Our relations are a solution to an issue we identify with ConceptNet, namely that there is overlap in meaning between different relations, and the majority of triples take a generic relation type.

Clarity in our ontology is further enforced by restricting the complexity of concept names to simple verb and noun phrases, ensuring that concepts are expressed at similar level of specificity. These phrasal concepts are stored as structured concepts, and we define structured relations to ensure that these remain well-connected to the graph and are easily surfaced when a model extracts a subgraph to operate on.

We translate the facts provided by WorldTree into MOntCS with a semi-automatic process and evaluate how useful it is for question answering. We identify that the graph-based model we use, QA-GNN, relies heavily on its masked language model component, which smooths over the differences in performance caused by different graphs. We find that MOntCS gives the highest accuracy amongst three other graphs on WorldTree question answering, and that a science-specific knowledge graph performs below random. We perform data ablation experiments to investigate the impact of different types of relation on question answering performance, and provide empirical evidence for suggestions in prior work that taxonomic relations are not useful for common sense question answering.

Our correlation analysis suggests that graph density and clustering coefficient correlate with performance.

References

[1] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. https://www.aclweb.org/anthology/N19-1423.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, Language Models are Few-Shot Learners, in: *Advances in Neural Information Processing Systems*, Vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan and H. Lin, eds, Curran Associates, Inc., 2020, pp. 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[3] P. Clark, O. Etzioni, T. Khot, B.D. Mishra, K. Richardson, A. Sabharwal, C. Schoenick, O. Tafjord, N. Tandon, S. Bhakthavatsalam, D. Groeneveld and M. Guerquin, From 'F' to 'A' on the N.Y. Regents Science Exams: An Overview of the Aristo Project, *arXiv 1909.01958* (2019). https://arxiv.org/abs/1909.01958.

[4] Y. Feng, X. Chen, B.Y. Lin, P. Wang, J. Yan and X. Ren, Scalable Multi-Hop Relational Reasoning for Knowledge-Aware Question Answering, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, 2020, pp. 1295–1309. doi:10.18653/v1/2020.emnlp-main.99. https://aclanthology.org/2020.emnlp-main.99.

[5] M. Yasunaga, H. Ren, A. Bosselut, P. Liang and J. Leskovec, QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering, in: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Online, 2021, pp. 535–546. doi:10.18653/v1/2021.naacl-main.45. https://aclanthology.org/2021.naacl-main.45.

[6] S. Ahn, H. Choi, T. Pärnamaa and Y. Bengio, A Neural Knowledge Language Model, *arXiv 1608.00318* (2016). https://arxiv.org/abs/1608.00318.

[7] B.Y. Lin, X. Chen, J. Chen and X. Ren, KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2829–2839. doi:10.18653/v1/D19-1282. https://aclanthology.org/D19-1282.

[8] H.P. Grice, Logic and Conversation, in: *Studies in the Way of Words*, P. Grice, ed., Harvard University Press, 1967, pp. 41–58, 36209.

[9] M. Sap, R.L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N.A. Smith and Y. Choi, ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning, in: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, AAAI Press, 2019, pp. 3027–3035. doi:10.1609/aaai.v33i01.33013027.

[10] R. Speer, J. Chin and C. Havasi, ConceptNet 5.5: An Open Multilingual Graph of General Knowledge, in: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, S.P. Singh and S. Markovitch, eds, AAAI Press, 2017, pp. 4444–4451. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972.

[11] N. Mostafazadeh, A. Kalyanpur, L. Moon, D. Buchanan, L. Berkowitz, O. Biran and J. Chu-Carroll, GLUCOSE: GeneraLized and COntextualized Story Explanations, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, 2020, pp. 4569–4586. doi:10.18653/v1/2020.emnlp-main.370. https://aclanthology.org/2020.emnlp-main.370.

[12] C. Fillmore, The Case for Case, in: *Universals in Linguistic Theory*, E. Bach and R.T. Harms, eds, Holt, Rinehart and Winston, New York, NY, USA, 1968, pp. 1–88.

[13] J. Gruber, Studies in Lexical Relations, Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1965.

[14] W.G. Lehnert, Plot Units and Narrative Summarization, *Cognitive Science* **5**(4) (1981), 293–331. doi:10.1207/s15516709cog0504_1. https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog0504_1.

[15] P. Jansen, N. Balasubramanian, M. Surdeanu and P. Clark, What's in an Explanation? Characterizing Knowledge and Inference Requirements for Elementary Science Exams, in: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 2956–2965. https://www.aclweb.org/anthology/C16-1278.

[16] P. Jansen, E. Wainwright, S. Marmorstein and C. Morrison, WorldTree: A Corpus of Explanation Graphs for Elementary Science Questions supporting Multi-hop Inference, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan, 2018. https://aclanthology.org/L18-1433.

[17] H. Liu and P. Singh, ConceptNet — A Practical Commonsense Reasoning Tool-Kit, *BT Technology Journal* **22**(4) (2004), 211–226. doi:10.1023/B:BTTJ.0000047600.45421.6d.

[18] C. Fellbaum, A Semantic Network of English: The Mother of All WordNets, *Computers and the Humanities* **32**(2–3) (1998), 209–220. doi:10.1023/A:1001181927857.

[19] T. Mihaylov and A. Frank, Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 821–832. doi:10.18653/v1/P18-1076. https://aclanthology.org/P18-1076.

[20] B. Dalvi Mishra, N. Tandon and P. Clark, Domain-Targeted, High Precision Knowledge Extraction, *Transactions of the Association for Computational Linguistics* **5** (2017), 233–246. doi:10.1162/tacl_a_00058. https://www.aclweb.org/anthology/Q17-1017.

[21] A. K M, S. Basu Roy Chowdhury and A. Dukkipati, Learning beyond Datasets: Knowledge Graph Augmented Neural Networks for Natural Language Processing, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 313–322. doi:10.18653/v1/N18-1029. https://aclanthology.org/N18-1029.

[22] D. Fried, P. Jansen, G. Hahn-Powell, M. Surdeanu and P. Clark, Higher-order Lexical Semantic Models for Non-factoid Answer Reranking, *Transactions of the Association for Computational Linguistics* **3** (2015), 197–210. doi:10.1162/tacl_a_00133. https://aclanthology.org/Q15-1015.

[23] C.F. Baker, C.J. Fillmore and J.B. Lowe, The Berkeley FrameNet Project, in: *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, 1998. https://aclanthology.org/C98-1013.

[24] C. Bonial, J. Bonn, K. Conger, J.D. Hwang and M. Palmer, PropBank: Semantics of New Predicate Types, in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 3013–3019. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1012_Paper.pdf.

[25] K.K. Schuler, Verbnet: A broad-coverage, comprehensive verb lexicon, Ph.D. Thesis, University of Pennsylvania, Philadelphia, Pennsylvania, USA, 2005.

[26] W.A. Woods, What's in a Link: Foundations for Semantic Networks, Technical Report, BOLT BERANEK AND NEWMAN INC CAMBRIDGE MA, 1975. https://apps.dtic.mil/sti/citations/ADA022584.

[27] D. Jurafsky and J.H. Martin, *Speech and Language Processing*, 3rd (draft) edn, USA, 2020.

[28] Princeton University, About WordNet, Princeton University, 2010.

[29] J. Lyons, *Semantics*, Vol. 2, Cambridge University Press, 1977. doi:10.1017/CBO9780511620614.

[30] M.L. Murphy (ed.), Hyponymy, meronymy, and other relations, in: *Semantic Relations and the Lexicon: Antonymy, Synonymy and other Paradigms*, Cambridge University Press, Cambridge, 2003, pp. 216–236. ISBN 978-0-521-78067-4. doi:10.1017/CBO9780511486494.007.

[31] Relations · commonsense/conceptnet5 Wiki. https://github.com/commonsense/conceptnet5.

[32] R.J. Brachman, What's in a concept: structural foundations for semantic networks, *International Journal of Man-Machine Studies* **9**(2) (1977), 127–152. doi:10.1016/S0020-7373(77)80017-5. https://www.sciencedirect.com/science/article/pii/S0020737377800175.

[33] S. Bird, E. Klein and E. Loper, *Natural Language Processing with Python*, O'Reilly Media, Inc., 2009. ISBN 978-0-596-51649-9.

[34] M. Honnibal, I. Montani, S. Van Landeghem and A. Boyd, spaCy: Industrial-strength Natural Language Processing in Python, Zenodo, 2020. doi:10.5281/zenodo.1212303.

[35] P. Pasupat and P. Liang, Compositional Semantic Parsing on Semi-Structured Tables, 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)*, Association for Computational Linguistics, Beijing, China, 2015, pp. 1470–1480. doi:10.3115/v1/P15-1142. https://aclanthology.org/P15-1142.

[36] T. Mihaylov, P. Clark, T. Khot and A. Sabharwal, Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering, in: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2018, pp. 2381–2391. http://aclweb.org/anthology/D18-1260.

[37] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao and J. Han, On the Variance of the Adaptive Learning Rate and Beyond, in: *International Conference on Learning Representations*, 2020. https://openreview.net/forum?id=rkgz2aEKDr.

[38] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, *arXiv 1907.11692* (2019). https://arxiv.org/abs/1907.11692.

[39] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest and A. Rush, Transformers: State-of-the-Art Natural Language Processing, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Online, 2020, pp. 38–45. doi:10.18653/v1/2020.emnlp-demos.6. https://aclanthology.org/2020.emnlp-demos.6.

[40] Y. Wu and K. He, Group Normalization, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[41] K. Richardson and A. Sabharwal, What Does My QA Model Know? Devising Controlled Probes Using Expert Knowledge, *Transactions of the Association for Computational Linguistics* **8** (2020), 572–588. doi:10.1162/tacl_a_00331. https://www.aclweb.org/anthology/2020.tacl-1.37.

[42] E. Davis and G. Marcus, Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence, *Communications of the ACM* **58**(9) (2015), 92–103. doi:10.1145/2701413. http://dl.acm.org/citation.cfm?doid=2817191.2701413.

[43] B. Tversky and K. Hemenway, Objects, parts, and categories, *Journal of Experimental Psychology: General* **113**(2) (1984), 169–193. doi:10.1037/0096-3445.113.2.169.

[44] J. Gordon and B. Van Durme, Reporting Bias and Knowledge Acquisition, in: *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, ACM, New York, NY, USA, 2013, pp. 25–30. ISBN 978-1-4503-2411-3. doi:10.1145/2509558.2509563.

[45] M.S. Schlichtkrull, N.D. Cao and I. Titov, Interpreting Graph Neural Networks for NLP With Differentiable Edge Masking, 2020. https://openreview.net/forum?id=WznmQa42ZAx.

[46] C.J. Fillmore, Frame Semantics and the Nature of Language, *Annals of the New York Academy of Sciences* **280**(1) (1976), 20–32. doi:10.1111/j.1749-6632.1976.tb25467.x.

[47] D. Vrandečić and M. Krötzsch, Wikidata: a free collaborative knowledgebase, *Communications of the ACM* **57**(10) (2014), 78–85. doi:10.1145/2629489.

[48] X. Wang, P. Kapanipathi, R. Musa, M. Yu, K. Talamadupula, I. Abdelaziz, M. Chang, A. Fokoue, B. Makni, N. Mattei and M. Witbrock, Improving Natural Language Inference Using External Knowledge in the Science Questions Domain, *Proceedings of the AAAI Conference on Artificial Intelligence* **33**(01) (2019), 7208–7215. doi:10.1609/aaai.v33i01.33017208. https://wvvw.aaai.org/ojs/index.php/AAAI/article/view/4705.

[49] W. Zhong, D. Tang, N. Duan, M. Zhou, J. Wang and J. Yin, Improving Question Answering by Commonsense-Based Pre-Training, *arXiv 1809.03568* (2018). https://arxiv.org/abs/1809.03568.

[50] B. Yang and T. Mitchell, Leveraging Knowledge Bases in LSTMs for Improving Machine Reading, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1436–1446. doi:10.18653/v1/P17-1132. https://aclanthology.org/P17-1132.

[51] R. Speer and C. Havasi, Representing General Relational Knowledge in ConceptNet 5, in: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 3679–3686. http://www.lrec-conf.org/proceedings/lrec2012/pdf/1072_Paper.pdf.

[52] D.L. McGuinness, R. Fikes, J. Rice and S. Wilder, An Environment for Merging and Testing Large Ontologies, in: *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning*, KR'00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000, pp. 483–493.

[53] Y. Zhou, S. Schockaert and J. Shah, Predicting ConceptNet Path Quality Using Crowdsourced Assessments of Naturalness, in: *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, L. Liu, R.W. White, A. Mantrach, F. Silvestri, J.J. McAuley, R. Baeza-Yates and L. Zia, eds, ACM, 2019, pp. 2460–2471. doi:10.1145/3308558.3313486.

[54] H. Rashkin, M. Sap, E. Allaway, N.A. Smith and Y. Choi, Event2Mind: Commonsense Inference on Events, Intents, and Reactions, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 463–473. doi:10.18653/v1/P18-1043. https://aclanthology.org/P18-1043.