Commonsense Knowledge in Wikidata

Filip Ilievski[1], Pedro Szekely[1], and Daniel Schwabe[2]

[1] Information Sciences Institute, University of Southern California
[2] Dept. of Informatics, Pontificia Universidade Católica Rio de Janeiro
{ilievski,pszekely}@isi.edu, dschwabe@inf.puc-rio.br

Abstract. Wikidata and Wikipedia have been proven useful for reasoning in natural language applications, like question answering or entity linking. Yet, no existing work has studied the potential of Wikidata for commonsense reasoning. This paper investigates whether Wikidata contains commonsense knowledge which is complementary to existing commonsense sources. Starting from a definition of common sense, we devise three guiding principles, and apply them to generate a commonsense subgraph of Wikidata (*Wikidata-CS*). Within our approach, we map the relations of Wikidata to ConceptNet, which we also leverage to integrate Wikidata-CS into an existing consolidated commonsense graph. Our experiments reveal that: 1) albeit Wikidata-CS represents a small portion of Wikidata, it is an indicator that Wikidata contains relevant commonsense knowledge, which can be mapped to 15 ConceptNet relations; 2) the overlap between Wikidata-CS and other commonsense sources is low, motivating the value of knowledge integration; 3) Wikidata-CS has been evolving over time at a slightly slower rate compared to the overall Wikidata, indicating a possible lack of focus on commonsense knowledge. Based on these findings, we propose three recommended actions to improve the coverage and quality of Wikidata-CS further.

Keywords: Commonsense Knowledge · Wikidata · Knowledge Graphs

1 Introduction

Common sense is "the basic ability to perceive, understand, and judge things that are shared by nearly all people and can be reasonably expected of nearly all people without need for debate" [10]. For instance, humans typically know that the political opposition is an opposite of the government, that hunger causes one to eat, and that if one walks in the rain one gets wet. Possessing such commonsense knowledge is important for both humans and machines in order to fill gaps in communication, and fulfill tasks such as entity recognition and linking from text, question answering, and planning. Yet, understanding common sense is difficult for machines. Even with the recent progress of language models such as BERT [6] and GPT-2 [22], which have been able to perform very well on a number of tasks with enough training,[1] the correct answer is often given for wrong reasons [8]. The utterances produced are syntactically sound, but may lack plausibility. For instance, GPT-2 complements the following prompt *'if you break a bottle that contains liquids, some of the liquid will (other things being equal) probably...'* with *'...wind up 300 meters away'* [18].

Commonsense graphs like ConceptNet [25] and ATOMIC [24] provide relevant knowledge that can be used to enhance the ability of language models to reason on downstream tasks. Unfortunately, these are largely incomplete, e.g., while ConceptNet contains information that a barbecue can be located in a garage, it is unable to infer that they are also common in other outdoor places, like parks, nor it has information on the expectations from such an event.

According to [27], common knowledge graphs (KGs) derived from Wikipedia, such as Wikidata [31] or YAGO4 [30], provide knowledge which is 'often required to achieve a deep understanding of both the low- and high-level concepts found in language'. In addition, Wikipedia has been used by a large number of systems for downstream reasoning tasks [11]. As the largest and highest-quality structured counterpart of Wikipedia [9], Wikidata is likely to contain useful commonsense knowledge - yet, no existing work has studied its commonsense coverage.

In this paper, we investigate whether Wikidata contains commonsense knowledge and whether that is complementary to existing commonsense knowledge graphs. Its contributions are: 1. we formulate three key principles for distinguishing commonsense knowledge from the rest in Wikidata, starting from three key properties of commonsense knowledge and from a survey of existing commonsense KGs (Section 3.1). These principles dictate that commonsense knowledge concerns well-known concepts and general-domain relations. 2. Based on them, we design and implement computational steps to extract a commonsense subgraph from Wikidata which we refer to as *Wikidata-CS* in the remainder of this paper (Section 3.2). Here, we also map relations in Wikidata to relations in ConceptNet. 3. We leverage this mapping to integrate Wikidata-CS into the Commonsense Knowledge Graph (CSKG) [13], which already contains well–known commonsense sources, such as ATOMIC and ConceptNet (Section 3.3). 4. We perform quantitative and qualitative analysis of the resulting subgraph (Section 4.1). Moreover, we compute overlaps between Wikidata-CS and other resources included in CSKG, like ConceptNet and WordNet (Section 4.2). 5. We perform the same experiments with three different versions of Wikidata from 2017, 2018, and 2020, and compare the results (Section 4.3). This allows us to quantify the evolution of commonsense knowledge in Wikidata over time. 6. In Section 5, we reflect on the findings from our experiments and propose recommended actions for further inclusion of commonsense knowledge in Wikidata.

2 Related Work

We review: 1. well-known commonsense KGs 2. prior works on reasoning with Wikidata or Wikipedia over text 3. studies of completeness of Wikidata.

[1] For example: https://leaderboard.allenai.org/socialiqa/submissions/public

Commonsense KGs such as ConceptNet and ATOMIC are popular and have been utilized by downstream reasoners [17]. Lexical resources, like WordNet [19] and FrameNet [1], capture commonsense knowledge about concepts and frames, respectively. Moreover, sources like Visual Genome [15] which have been originally proposed for a different purpose (image captioning and visual recognition), have recently been recognized as sources of commonsense knowledge. Commonsense knowledge can also be extracted from documents [29, 20], query logs [23], or quantities [7]. A recent resource, called the Commonsense Knowledge Graph [13], consolidates many of these resources into a single KG. The complementarity of these sources motivates their integration, but also reveals that they are still largely incomplete. Wikidata, as one of the richest public KGs, holds a promise to enrich the set of recorded commonsense facts even further.

A recent idea is to use language models, like BERT [6] and GPT-2 [22], as knowledge bases, due to their inherent ability to produce a fact for any input prompt. Still, they often exhibit shallow understanding of the world [8]. Integration with KGs like Wikidata or ConceptNet may increase their robustness [17].

Reasoning with Wikipedia and Wikidata Wikipedia and Wikidata serve as sources of background knowledge in natural language processing tasks [11], e.g., as a repository of entities to link to, or as a source of contextual information to help linking entities in text [21, 3, 5]. The work by Suh et al. [28] attempts to extract commonsense knowledge from Wikipedia. As far as we are aware, there is no comprehensive proposal to extract commonsense knowledge from Wikipedia or Wikidata, or to study their strengths and weaknesses for this purpose.

Studies of completeness of Wikidata Several papers study the completeness of Wikidata [4, 2, 16]. Luggen et al. [16] provide an approach to estimate class completeness in knowledge graphs, and use Wikidata as a use case. They note that some classes in Wikidata, like Painting, are more complete than others, such as Mountain. In addition, they also quantify the evolution of Wikidata over time. Similarly, we also study the completeness of Wikidata and its richness over time, albeit focusing on its coverage of commonsense knowledge.

3 Extraction of Commonsense Knowledge from Wikidata

3.1 Principles of commonsense knowledge

Common sense is "the basic ability to perceive, understand, and judge things that are shared by nearly all people and can be reasonably expected of nearly all people without need for debate" [10]. From this definition, we can infer that commonsense knowledge: 1) concerns conceptual rather than instance-based information; 2) is primarily about commonly known observations; 3) targets general-domain information. We expand these three aspects into three guiding principles for our approach, which would allow us to define a commonsense subset of the knowledge in a general KG such as Wikidata.

P1 **Concepts, not entities** The primary principle of commonsense knowledge draws on the distinction between concept- and named-entity-level (instance) knowledge. Generally speaking, most concept-level knowledge is common sense, whereas most named-entity-level knowledge is not. The fact that *houses* have *rooms* is commonsense knowledge, as it common and widely applicable; the fact that the *Versailles Palace* has *700 rooms* is not, as it concerns a particular instance and cannot be expected by most people. Thus, principle P1 is that commonsense knowledge has to be about concepts.

P2 **Commonness** The second principle (P2) of commonsense knowledge is its 'commonness': it is knowledge about well-known concepts that is shared among most human beings. The fact that a *container* (Q987767) is used for *storage* (Q9158768) is a common fact, whereas the fact that *noma* (Q994794) is a subclass of *aphthous stomatitis* (Q189956) is fairly unknown.

P3 **General-domain knowledge** The third principle is that commonsense knowledge is about general-domain information rather than expert knowledge about a specific domain like chemistry or biology. Notably, even within a knowledge type, some relations describe general information, whereas others require expert knowledge. For instance, considering meronymy relations, we observe that *part of* describes well-known facts (e.g., wheel is part of a car), whereas *cell component* focuses on biological knowledge (e.g., *cholesterol* has component *cell membrane*). As a third principle (P3), we aim to distinguish between domain-specific and general-domain knowledge.

3.2 Approach

Next, we apply P1-P3 to select commonsense knowledge from Wikidata.

1. Excluding named entities (P1) In practice, Wikidata does not make a clear distinction between concepts and named instances through its structured knowledge. The relation *instance of* (`P31`) would intuitively be useful for this; yet, it often expresses an *is-a* relation between concepts, similar to *subclass of* (`P279`). For instance, Wikidata states that *surgeon* is an instance of *medical profession*, and a subclass of *medical specialist*. Leveraging the `rdf:type` relation from another public ontology, such as DBpedia, is a possible direction, yet this strategy would be limited to the set of nodes that are mapped between Wikidata and DBpedia. Hence, we follow a different route. The convention of Wikidata stipulates that the labels of named entities should be capitalized, whereas the ones for concepts should not.[2] Following this rule, we employ a simple heuristic of selecting edges where both nodes have alphanumeric labels starting with a lowercase letter. We expand this rule and filter out labels that contain any capital letter, to remove entities with labels like "graf Nikolai Aleksyeevich Sheremetev". This procedure implicitly excludes nodes without English labels.

2. Characterizing commonness (P2) We argued that commonsense facts concern common concepts. Wikidata-based metrics of frequency or popularity, such as PageRank, cannot be used to estimate commonness, as they inherit the bias towards topics that are heavily represented in Wikidata (e.g., entertainment or science). Instead, we approximate commonness by frequencies of word and

[2] https://www.wikidata.org/wiki/Help:Label

Table 1. Number of edges and representative examples for the top 20 relations.

Relation	#edges	Examples
subclass of (P279)	172,535	saxophone - woodwind instrument
instance of (P31)	141,499	happiness - positive emotion
part of (P361)	9,118	shower - bathroom
different from (P1889)	7,767	vein - artery
has part (P527)	6,252	senses - touch
cell component (P681)	5,607	cholesterol - cell membrane
property constraint (P2302)	5,180	votes received - integer constraint
facet of (P1269)	4,792	wind - weather
strand orientation (P2548)	4,345	sac-1 - forward strand
use (P366)	3,045	crystal ball - psychic reading
opposite of (P461)	3,028	political opposition - government
properties for this type (P1963)	2,382	human - date of birth
molecular function (P680)	2,369	protein kinase - kinase activity
see also (P1659)	2,344	position held - member of
sport (P641)	2,338	head stand - gymnastics
followed by (P156)	2,244	middle school - secondary school
follows (P155)	2,234	queen - jack
material used (P186)	2,047	ice cream cone - wafer
is a list of (P360)	1,914	list of major opera composers - human
Wikidata property (P1687)	1,746	president - head of government

phrase usage that have been pre-computed over an independent corpus [26].[3] Here, we assume that frequently occurring words and phrases refer to well-known concepts. According to this tool, the frequency of a common word, like *storage*, is much higher than that of a relatively unknown word, such as *noma* (3.39e-05 compared to 3.24e-07). We select edges where both the subject and the object labels have usage frequency above an empirically determined threshold of $1e-06$.

3. Excluding domain knowledge (P3) The initial two steps yield 420,822 edges, involving 414 edge types and describing 194,595 nodes. Table 1 presents the number of occurrences for the 20 most frequent edge types, together with a representative example edge for each type. By analyzing the frequency distribution of the remaining relations, we observe that the frequency quickly decays. The 50th most common relation describes less than 500 edges, and their frequency plot becomes relatively flat (Figure 1). Hence, we focus on the 50 most frequent relations and distinguish the remaining knowledge by manually mapping them to relations in ConceptNet v5.7.[4] These account for 409,775 edges, which is 97.4% of the total set of edges available at this point.[5]

The main guideline for this mapping was to exclude properties which are meant to describe domain-specific information, such as *strand orientation* (`P2548`).

[3] https://pypi.org/project/wordfreq/
[4] https://github.com/commonsense/conceptnet5/wiki/Downloads
[5] In the future, we intend to consolidate the remaining statements of Wikidata by mapping them to ConceptNet relations as well.

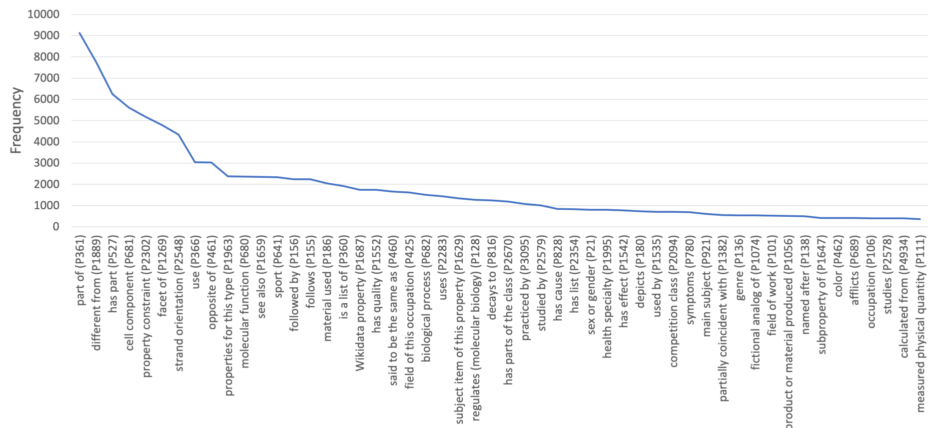

Fig. 1. Frequency distribution of the 50 most frequent remaining relations. For readability, we exclude P31 and P279, as these are much more populated than the rest.

The mapping was performed independently by two authors of this paper. In all cases, the annotators agreed on whether the relation describes general-domain knowledge. In 9 cases, the annotators disagreed on which ConceptNet relation is the most appropriate to map to. Typically, this meant that ConceptNet lacks a relation with the same specificity, forcing the annotators to opt for a more generic relation, such as `/r/HasContext`. The disagreements were resolved through a joint discussion and examination of exemplar edges in Wikidata.

The resulting mappings are shown in Table 2. 44 out of the top 50 relations were mapped to existing relations in ConceptNet, yielding 388,250 edges. The remaining six relations are either biology domain-specific: cell component (P681), strand orientation (P2548), molecular function (P680), biological process (P682); physical domain-specific: decays to (P816); or ontological: property constraint (P2302). The mapping shows that some ConceptNet properties (e.g., `/r/Antonym`) have a single counterpart in Wikidata (*opposite of*), while others (e.g., `/r/HasContext`) map to several properties, often with more specific meanings (e.g., *genre*, *sport*). This might reveal an opportunity to enrich the specificity of relations in ConceptNet with more detailed ones as in Wikidata.[6] Some relations in ConceptNet (e.g., `/r/MotivatedByGoal`) may have no counterpart in Wikidata, and others map to a relation which is very sparse for common concepts. For instance, `/r/AtLocation` maps to *location*, which is well-populated for named entities in Wikidata, but only ranks 72nd with 159 occurrences in our commonsense subset. These observations reveal a knowledge gap in Wikidata. In several cases, the relation in Wikidata is inverse to that in ConceptNet, e.g., *has part* to `/r/PartOf` and *has cause* to `/r/Causes`. We analyze the overlap between ConceptNet and Wikidata further in Section 4.2.

[6] For some ConceptNet relations, like `/r/PartOf` and `/r/HasProperty`, a similar proposal to add detail comes from WebChild [29].

Table 2. Mapping of relations between Wikidata and ConceptNet. The Wikidata relations prefixed with '*' are inverse to the relation in ConceptNet.

Category	**ConceptNet**	**Wikidata**
distinctness	/r/DistinctFrom	different from (P1889)
antonymy	/r/Antonym	opposite of (P461)
synonymy	/r/Synonym	said to be the same as (P460)
similarity	/r/SimilarTo	partially coincident with (P1382)
derivation	/r/DerivedFrom	named after (P138), fictional analog of (P1074)
inheritance	/r/IsA	instance of (P31), subclass of (P279), subproperty of (P1647)
meronymy	/r/PartOf	part of (P361), *has part (P527), *has parts of the class (P2670)
material	/r/MadeOf	material used (P186), is a list of (P360), *has list (P2354)
attribution	/r/CreatedBy	*product or material produced (P1056)
utility	/r/UsedFor	use (P366), *uses (P2283), used by (P1535)
properties	/r/HasProperty	color (P462), has quality (P1552), properties of this type (P1963), Wikidata property (P1687), sex or gender (P21)
causation	/r/Causes	*has cause (P828), has effect (P1542), symptoms (P780)
ordering	/r/HasPrerequisite	*followed by (P156), follows (P155)
context	/r/HasContext	facet of (P1269), field of this occupation (P425), health specialty (P1995), main subject (P921), competition class (P2094), genre (P136), studied by (P2579), field of work (P101), afflicts (P689), *practiced by (P3095), depicts (P180), sport (P641)
other	/r/RelatedTo	see also (P1659), subject item of this property (P1629)

Finally, assuming that domain-specific relations involve domain-specific nodes, we construct a set of 'blacklist' nodes found in these relations. We ensure that the remaining edges do not contain these domain-specific nodes. This allows us to filter out nodes like protein (`Q8054`), which has over 172 thousand incoming edges, typically from child proteins.

3.3 Integration in the Commonsense Knowledge Graph

The Commonsense Knowledge Graph (CSKG) [13] is an existing resource that consolidates information from seven commonsense sources, including ConceptNet, Roget [14], Visual Genome [15], WordNet [19], and Wikidata. It is represented using the KGTK [12] format with 10 columns, including the core elements of an edge (`id`, `node1`, `relation`, and `node2`), their labels (e.g., `node1;label`, and provenance information about an edge (`source` and `sentence`). Regarding Wikidata, CSKG includes all the edges involving the inheritance (`P279`) relation.

We integrate the commonsense subset of Wikidata presented in this paper into CSKG. For this purpose, we adapt its columns to match those specified by CSKG. The columns for which we lack information, such as `sentence`, are left empty. We map the 50 most frequent relations to ConceptNet relations following Table 2, and discard the small number of remaining statements.

3.4 Implementation

We implement the proposed selection of commonsense knowledge from Wikidata by using the Knowledge Graph ToolKit (KGTK) [12]. KGTK allows us to carry out the proposed approach in a direct and simple way, despite the challenging size and complexity of Wikidata. The full experiment reported in this paper is coded as three Jupyter Notebooks which run on a laptop in under an hour.[7] The starting point is the entire Wikidata split into three Wikidata files in KGTK tabular format (an edge file, a node file, and a qualifiers file), as pre-computed with the *import-wikidata* command.

The concrete steps are as follows. We use a customized Python function to create a subset of the node file that contains only concept nodes, by removing nodes whose labels are either empty or contain a capital letter. We use the *ifexists* join operator to filter out edges that do not connect two concepts from the edge file.The command *remove-columns* trims all columns which are not necessary for the experiment. After this, we run *compact* to remove duplicate edges. At this point, we have a subset of edges that are about concepts (P1). To prepare for the usage filtering and help human readability, we expand the set of columns with the *lift* command to include the labels of the subject, the object, and the relation. We use the aforementioned threshold-based filter to select edges for which both the subject and the object are common concepts. Next, we inspect the remaining edges in terms of their relations. We apply the manual mapping of the top 50 relations (Section 3.2) to consolidate the remaining Wikidata graph and make its edge types compatible with the format of CSKG.

These steps produce the subset of Wikidata (*Wikidata-CS*), which satisfies our principles (P1-P3), in the CSKG format. Finally, we use *graph-statistics* to compute metrics over this subset. Wikidata-CS is available for download.[8]

4 Analysis

4.1 General Statistics

Wikidata-CS consists of 71,243 nodes and 106,103 edges. It uses 44 edge types to describe these edges. The mean node degree is 2.98, which is higher than in the *subclass of* subset of Wikidata (2.45) [13]. The nodes with the highest PageRank in the resulting graph are: *artificial entity* (`Q16686448`), *kinship* (`Q171318`), and *class* (`Q16889133`), which are more customary compared to the top nodes in the unfiltered subclass-of data, all of which describe biochemical concepts [13].

[7] https://github.com/usc-isi-i2/cskg/tree/master/wikidata
[8] https://doi.org/10.5281/zenodo.3983029

Table 3. Comparison of the size of Wikidata-CS to commonsense sources in CSKG.

	ATOMIC	**Concept Net**	**Frame Net**	**Roget**	**Visual Genome**	**Word Net**	**Wikidata-CS (this paper)**
# nodes	304,909	1,787,373	36,582	71,804	10,830	91,294	**71,243**
# edges	732,723	3,423,004	79,060	1,403,461	2,218,868	111,276	**106,103**

The five most frequent relations in Wikidata-CS are: *subclass of* (`P279`), *instance of* (`P31`), *different from* (`P1889`), *part of* (`P361`), and *has part* (`P527`). The first two account for 68.8% of all edges, indicating the the commonsense knowledge in Wikidata mostly concerns taxonomic information. After mapping the relations to ConceptNet, all commonsense knowledge corresponds to 15 edge types. We perform de-duplication to consolidate edges that were expressed with relations of the same group (e.g., *subclass of* and *instance of*), or in two directions with inverse properties (e.g., *has cause* and *has effect*). The distribution of knowledge across these types is shown in Table 4 (last column). The final set has 101,771 edges, which is below 0.01% of the full Wikidata. Next, we compare the content and size of Wikidata-CS to those of other commonsense KGs.

4.2 Comparison to other graphs in CSKG

The integration of Wikidata-CS into CSKG allows one to easily compare its content to other sources, such as ConceptNet. How does the size of the commonsense subset in Wikidata compare to that of the other sources? How much of the commonsense knowledge in Wikidata is already present in these other sources? How much is missing? Conversely: how many edges are defined in ConceptNet or WordNet, but are lacking in Wikidata? We provide insight into these questions.

Table 3 compares the size of Wikidata-CS with the other subgraphs within CSKG. Despite the fact that Wikidata is by far the largest graph, its commonsense subset ranks 6th in terms of edges and 5rd in terms of nodes, being only larger than FrameNet and over 30 times smaller than ConceptNet.[9] We also inspect the overlap between the knowledge in Wikidata-CS and in other CSKG sources that share the same relations. Since only the relations are mapped between these sources, whereas the nodes are not, we assume equivalence of two edges with identical subject labels, object labels, and edge types. The results are given in Table 5. We observe that Wikidata-CS shares 2,386 edges with ConceptNet, 1,613 with WordNet, and only 299 with Roget. Above all, this investigation shows extremely little overlap between Wikidata-CS and the other three graphs. The observation that commonsense knowledge in Wikidata is almost entirely missing in the other KGs, and vice versa, validates the main pursuit of this paper, and motivates the consolidation of these sources into a single graph.

[9] To be fair, the edge count of the other graphs may include edges with named entities (e.g., through the `/r/IsA` relation), which were excluded in Wikidata-CS.

Table 4. Temporal evolution of the Wikidata commonsense knowledge.

	2017-12-27	2018-12-10	2020-05-04
/r/IsA	31,668	45,606 (144%)	72,707 (230%)
/r/PartOf	3,390	4,416 (130%)	7,938 (234%)
/r/HasContext	1,968	3,189 (162%)	6,152 (313%)
/r/DistinctFrom	782	2,011 (257%)	4,934 (631%)
/r/HasPrerequisite	413	1,965 (476%)	4,131 (1,000%)
/r/UsedFor	735	1,215 (165%)	2,469 (336%)
/r/Antonym	1,109	1,530 (138%)	2,184 (197%)
/r/MadeOf	415	834 (201%)	1,426 (344%)
/r/Synonym	478	655 (137%)	1,070 (224%)
/r/HasProperty	339	650 (192%)	1,049 (309%)
/r/Causes	150	238 (159%)	651 (434%)
/r/DerivedFrom	190	293 (154%)	540 (284%)
/r/SimilarTo	28	77 (275%)	345 (1,232%)
/r/CreatedBy	51	68 (133%)	187 (367%)
/r/RelatedTo	33	40 (121%)	42 (127%)
edges (Wikidata-CS)	41,769	62,787 (150%)	101,771 (244%)
edges (Wikidata)	405,081,219	696,605,955 (172%)	1,105,944,515 (273%)
nodes (Wikidata-CS)	32,620	47,056	71,243
nodes (Wikidata)	42,187,222	53,004,762	84,601,621

Table 5. Overlap between Wikidata and other commonsense knowledge sources.

Other source	Both	Wikidata-CS only	Other source only
ConceptNet	2,386	97,473 (97.6%)	3,320,935 (99.9%)
Roget	299	99,560 (99.7%)	1,403,162 (99.9%)
WordNet	1,613	98,246 (98.4%)	419,103 (99.6%)

We note that, with this lexical overlap approach, an edge might be counted multiple times if its nodes have multiple labels. This is why WordNet has over 500k edges in total in Table 5, while having little over 100k in the original data. Future work should investigate more semantic overlap estimation methods.

4.3 Evolution of the Wikidata commonsense knowledge over time

The size of Wikidata has been growing at a tremendous rate. In only 30 months, its number of edges nearly tripled and the number of nodes doubled (Table 4). A natural question arises: has the size of its commonsense subset been growing at a similar rate? To investigate this question, we consider three versions of Wikidata, with dates: 2017-12-27, 2018-12-10, and 2020-05-04. For fair comparison, we apply our approach (Section 3.2) on the three Wikidata dumps.

For each of the Wikidata dumps, we present the number of edges per relation in Table 4. Firstly, while the number of edges in Wikidata-CS has multiplied for nearly all relations (except `RelatedTo`), its growth is slightly slower than the full Wikidata - 244% vs 273% between December 2017 and May 2020. A similar trend holds for the December 2018 version. Hence, despite the apparent interest in enriching the commonsense knowledge subset of Wikidata, this has not been a priority so far. Secondly, we see larger growth of the relations *SimilarTo*, *HasPrerequisite*, and *DistinctFrom* relative to the others. This shows that certain commonsense aspects (like differentiating potentially confusing concepts) may be more relevant to the Wikidata community and its applications than others.

5 Discussion

The commonsense knowledge in Wikidata could benefit applications like question answering or entity linking. For instance, let's consider the following true/false question from the CycIC dataset:[10] *Suppose something is under the table. It is either a toaster or a correction tape dispenser. You can tell that it isn't a kitchen tool. True or False: The thing under the table is a correction tape dispenser.* The key implicit knowledge in this example is the fact that the toaster is often found in the kitchen, while the dispenser is not. Luckily, over 100k such commonsense facts are part of our Wikidata-CS collection, and could help a downstream system to reason over such questions.

Still, we noted that only a neglegible portion of Wikidata directly describes commonsense knowledge today. Given the considerable community involved in Wikidata and Wikipedia, and the commonsense relations identified in this paper, we propose for the commonsense knowledge in Wikidata to be substantially enriched in the near future. We discuss three actions towards this goal:

1. Integration of ready commonsense sources into Wikidata A number of commonsense sources, like ConceptNet and ATOMIC, contain much complementary knowledge that could be included into Wikidata (cf. Table 3). Our prior work on consolidating their formats and modeling principles into CSKG enables their seamless integration into Wikidata, when so desired. At present, CSKG contains 5.89 million edges, expressed through 58 relations. The mappings in Table 2 could be used as a starting point, whereas missing relations might need to be added to Wikidata. Data licensing may be a a roadblock here.

2. Generalizing over instance-level knowledge Much of the commonsense knowledge in Wikidata is indirectly expressed through its instance-level knowledge. While Barack Obama being born in Hawaii is not a commonsense fact, the fact that humans have a birthplace is. Furthermore, all humans have a single birthplace, i.e., it is a functional property. One could think of other generalizations as well, e.g., if many locations belong to countries, it is common sense thinking that any location would belong to a country. Such commonsense information is not directly represented in Wikidata, yet it could be inferred by statistical generalization over instance-level knowledge.

3. Missing knowledge types The Wikidata model defines the notion of qualifiers, which would be ideal to represent much commonsense knowledge.

[10] https://leaderboard.allenai.org/cycic/submissions/get-started

However, in many cases, Wikidata does not only lack a commonsense fact, but also the relation or the qualifier that would express it. For instance, while many qualifiers (e.g., minimum and maximum value) express quantities, no qualifier describes typical/expected quantity.[11] This could express that spiders typically have eight legs, while chairs have four. Qualifiers for expressing a purpose or a goal (e.g., one participates in a competition in order to win) are also missing. Besides qualifiers, it might be of use to include relations that are currently missing, like typical properties of concepts (e.g., elephants are heavy), or their symbolism (e.g., red is a symbol of danger). The actual information could be extracted from unstructured sources, like Wikipedia, or reused from previous extractions [7].

6 Conclusion

Wikidata has been growing tremendously in terms of both size and popularity. Consequently, it has been attracting interest from applications that require background knowledge in order to fill in gaps, such as question answering and entity linking. In this paper, we studied the commonsense knowledge coverage of Wikidata and its complementarity to existing commonsense graphs. Starting from three key principles of commonsense, we devised a three-step filtering approach that distinguishes concepts from named entities, favors common concepts, and general-domain knowledge types. Here, we also created mappings between the relations in the commonsense subset of Wikidata (Wikidata-CS) and those in ConceptNet, which allowed us to integrate Wikidata-CS into CSKG, an existing consolidated graph of commonsense knowledge. We analyzed the content of Wikidata-CS and compared it to other existing sources, like ConceptNet and WordNet, noting that while Wikidata contains useful and novel commonsense knowledge that complements other sources, its coverage of commonsense knowledge is currently largely incomplete. We propose three directions to improve this in the future: by inclusion of the knowledge from the Commonsense Knowledge Graph, by generalizing over existing instance-level knowledge in Wikidata, and by inclusion of missing knowledge types that are relevant for representing commonsense knowledge. In addition, subsequent research should evaluate the quality of Wikidata-CS and its relevance for commonsense reasoning, based on user studies and downstream tasks.

Acknowledgements

We would like to thank *Daniel Garijo* for the very helpful comments and suggestions on drafts of this paper. We would also like to thank *Craig Milo Rogers* for his help with KGTK. We are grateful for the reviewer comments. This material is based upon work sponsored by the DARPA MCS program under Contract No. N660011924033 with the United States Office Of Naval Research and by Air Force Research Laboratory under agreement number FA8750-20-2-10002.

[11] https://www.wikidata.org/wiki/Wikidata:List_of_properties/Wikidata_qualifier

References

1. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley framenet project. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1. pp. 86–90 (1998)
2. Balaraman, V., Razniewski, S., Nutt, W.: Recoin: relative completeness in wikidata. In: Companion Proceedings of the The Web Conference 2018. pp. 1787–1792 (2018)
3. Cetoli, A., Bragaglia, S., O'Harney, A.D., Sloan, M., Akbari, M.: A neural approach to entity linking on wikidata. In: European Conference on Information Retrieval. pp. 78–86. Springer (2019)
4. Darari, F., Prasojo, E., Razniewski, S., Nutt, W.: Cool-wd: A completeness tool for wikidata. CEUR-WS. org (2017)
5. Delpeuch, A.: Opentapioca: Lightweight entity linking for wikidata. arXiv preprint arXiv:1904.09131 (2019)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Elazar, Y., Mahabal, A., Ramachandran, D., Bedrax-Weiss, T., Roth, D.: How large are lions? inducing distributions over quantitative attributes. arXiv preprint arXiv:1906.01327 (2019)
8. Ettinger, A.: What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the Association for Computational Linguistics **8**, 34–48 (2020)
9. Färber, M., Ell, B., Menne, C., Rettinger, A.: A comparative survey of dbpedia, freebase, opencyc, wikidata, and yago. Semantic Web Journal **1**(1), 1–5 (2015)
10. Gunning, D.: Machine common sense concept paper. arXiv preprint arXiv:1810.07528 (2018)
11. Hovy, E., Navigli, R., Ponzetto, S.P.: Collaboratively built semi-structured content and artificial intelligence: The story so far. Artificial Intelligence **194**, 2–27 (2013)
12. Ilievski, F., Garijo, D., Chalupsky, H., Divvala, N.T., Yao, Y., Rogers, C., Li, R., Liu, J., Singh, A., Schwabe, D., et al.: Kgtk: A toolkit for large knowledge graph manipulation and analysis. arXiv preprint arXiv:2006.00088 (2020)
13. Ilievski, F., Szekely, P., Cheng, J., Zhang, F., Qasemi, E.: Consolidating commonsense knowledge. arXiv preprint arXiv:2006.06114 (2020)
14. Kipfer, B.: Roget's 21st century thesaurus in dictionary form (éd. 3). new york: The philip lief group (2005)
15. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)
16. Luggen, M., Difallah, D., Sarasua, C., Demartini, G., Cudré-Mauroux, P.: Nonparametric class completeness estimators for collaborative knowledge graphs—the case of wikidata. In: International Semantic Web Conference. pp. 453–469. Springer (2019)
17. Ma, K., Francis, J., Lu, Q., Nyberg, E., Oltramari, A.: Towards generalizable neuro-symbolic systems for commonsense question answering. arXiv preprint arXiv:1910.14087 (2019)
18. Marcus, G.: The next decade in ai: four steps towards robust artificial intelligence. arXiv preprint arXiv:2002.06177 (2020)

19. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
20. Mishra, B.D., Tandon, N., Clark, P.: Domain-targeted, high precision knowledge extraction. Transactions of the Association for Computational Linguistics **5**, 233–246 (2017)
21. Mulang, I.O., Singh, K., Vyas, A., Shekarpour, S., Sakor, A., Vidal, M.E., Auer, S., Lehmann, J.: Context-aware entity linking with attentive neural networks on wikidata knowledge graph. arXiv preprint arXiv:1912.06214 (2019)
22. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)
23. Romero, J., Razniewski, S., Pal, K., Z. Pan, J., Sakhadeo, A., Weikum, G.: Commonsense properties from query logs and question answering forums. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. pp. 1411–1420 (2019)
24. Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N.A., Choi, Y.: Atomic: An atlas of machine commonsense for if-then reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 3027–3035 (2019)
25. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: An open multilingual graph of general knowledge. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
26. Speer, R., Chin, J., Lin, A., Jewett, S., Nathan, L.: Luminosoinsight/wordfreq: v2.2 (Oct 2018). https://doi.org/10.5281/zenodo.1443582, https://doi.org/10.5281/zenodo.1443582
27. Storks, S., Gao, Q., Chai, J.Y.: Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. arXiv preprint arXiv:1904.01172 pp. 1–60 (2019)
28. Suh, S., Halpin, H., Klein, E.: Extracting common sense knowledge from wikipedia. In: Proceedings of the Workshop on Web Content Mining with Human Language Technologies at ISWC. vol. 6 (2006)
29. Tandon, N., De Melo, G., Weikum, G.: Webchild 2.0: Fine-grained commonsense knowledge distillation. In: Proceedings of ACL 2017, System Demonstrations. pp. 115–120 (2017)
30. Tanon, T.P., Weikum, G., Suchanek, F.: Yago 4: A reason-able knowledge base. In: European Semantic Web Conference. pp. 583–596. Springer (2020)
31. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)