CSKG: The CommonSense Knowledge Graph

Filip Ilievski, Pedro Szekely, and Bin Zhang

Information Sciences Institute, University of Southern California
{iilievski,pszekely,binzhang}@isi.edu

Abstract. Sources of commonsense knowledge aim to support applications in natural language understanding, computer vision, and knowledge graphs. These sources contain complementary knowledge to each other, which makes their integration desired. Yet, such integration is not trivial because of their different foci, modeling approaches, and sparse overlap. In this paper, we propose to consolidate commonsense knowledge by following five principles. We apply these principles to combine seven key sources into a first integrated CommonSense Knowledge Graph (CSKG). We perform analysis of CSKG and its various text and graph embeddings, showing that CSKG is a well-connected graph and that its embeddings provide a useful entry point to the graph. Moreover, we show the impact of CSKG as a source for reasoning evidence retrieval, and for pre-training language models for generalizable downstream reasoning. CSKG and all its embeddings are made publicly available to support further research on commonsense knowledge integration and reasoning.

Resource type: Knowledge graph
License: CC BY-SA 4.0
DOI: https://doi.org/10.5281/zenodo.4331372
Repository: https://github.com/usc-isi-i2/cskg

Keywords: commonsense knowledge · knowledge graph · embeddings

1 Introduction

Recent commonsense reasoning benchmarks [25,3] and neural advancements [16,15] shed a new light on the longstanding task of capturing, representing, and reasoning over commonsense knowledge. While state-of-the-art language models [8,16] capture linguistic patterns that allow them to perform well on commonsense reasoning tasks after fine-tuning, their robustness and explainability could benefit from integration with structured knowledge, as shown by KagNet [15] and HyKAS [17]. Let us consider an example task question from the SWAG dataset [36], which describes a woman that takes a sit at the piano:

Q: On stage, a woman takes a seat at the piano. She:
1. sits on a bench as her sister plays with the doll.
2. smiles with someone as the music plays.
3. is in the crowd, watching the dancers.
4. nervously sets her fingers on the keys.
Answering this question requires knowledge that humans possess and apply, but machines cannot distill directly in communication. Luckily, graphs of (commonsense) knowledge contain such knowledge. ConceptNet’s [27] triples state that pianos have keys and are used to perform music, which supports the correct option and discourages answer 2. WordNet [20] states specifically, though in natural language, that pianos are played by pressing keys. According to an image description in Visual Genome, a person could play piano while sitting and having their hands on the keyboard. In natural language, ATOMIC [24] indicates that before a person plays piano, they need to sit at it, be on stage, and reach for the keys. ATOMIC also lists strong feelings associated with playing piano. FrameNet’s [1] frame of a performance contains two separate roles for the performer and the audience, meaning that these two are distinct entities, which can be seen as evidence against answer 3.

While these sources clearly provide complementary knowledge that can help commonsense reasoning, their different foci, representation formats, and sparse overlap makes integration difficult. Taxonomies, like WordNet, organize conceptual knowledge into a hierarchy of classes. An independent ontology, coupled with rich instance-level knowledge, is provided by Wikidata [32], a structured counterpart to Wikipedia. FrameNet, on the other hand, defines an orthogonal structure of frames and roles; each of which can be filled with a WordNet/Wikidata class or instance. Sources like ConceptNet or WebChild [29], provide more ‘episodic’ commonsense knowledge, whereas ATOMIC captures pre- and post-situations for an event. Image description datasets, like Visual Genome [13], contain visual commonsense knowledge. While links between these sources exist (mostly through WordNet synsets), the majority of their nodes and edges are disjoint.

In this paper, we propose an approach for integrating these (and more sources) into a single Common Sense Knowledge Graph (CSKG). We survey existing sources of commonsense knowledge to understand their particularities and we summarize the key challenges on the road to their integration (section 2). Next, we devise five principles and a representation model for a consolidated CSKG (section 3). We apply our approach to build the first version of CSKG, by combining seven complementary, yet disjoint, sources. We compute several graph and text embeddings to facilitate reasoning over the graph. In section 4, we analyze the content of the graph and the generated embeddings. We provide insights into the utility of CSKG for downstream reasoning on commonsense (QA) tasks in section 5. In section 6 we reflect on the learned lessons and list the next steps for CSKG. We conclude in section 7.

2 Problem statement

2.1 Sources of Common Sense Knowledge

Table 1 summarizes the content, creation method, size, external mappings, and example resources for representative public commonsense sources: ConceptNet [27], WebChild [29], ATOMIC [24], Wikidata [32], WordNet [20], Roget [12], VerbNet [26], FrameNet [1], Visual Genome [13], and ImageNet [7]. Primarily, we
Table 1. Survey of existing sources of commonsense knowledge.

| ConceptNet | everyday objects, actions, states, relations (multilingual) | 36 relations, 8M WordNet, 21M edges | /c/en/piano, /c/en/piano/n, /c/en/piano/n/wn | describe, creation, size, mappings, examples |
|-----------|-----------------------------------------------------------------|-------------------------------------|---------------------------------------------|-------------------------------------------|
| Web Child | everyday objects, actions, automatic states, relations extraction | 4 relation groups, 2M WordNet nodes, 18M edges | hasTaste, fasterThan | |
| ATOMIC    | event pre/post-conditions sourcing                              | 9 relations, 300k ConceptNet, Cyc, wanted-to impressed | wd:Q1234 wdt:P31 | |
| Wikidata  | instances, concepts, relations sourcing                         | 1.2k relations, 75M various objects, 900M edges | truncate, antonym | |
| WordNet   | words, concepts, manual relations                               | 10 relations, 155k words, 176k synsets, 877k edges | dog.n.01, hypernymy | |
| Roget     | words, relations manual                                         | 2 relations, 72k words, 1.4M edges | performance-v, performance-26.7-1 | |
| VerbNet   | verbs, relations manual                                         | 273 top classes 23 FrameNet, roles, 5.3k senses WordNet | Activity, Change-of_leadership New_leader | |
| FrameNet  | frames, roles, relations manual                                 | 1.9k edges, 1.2k frames, 12k roles, 13k lexical units | fire hydrant, white dog | |
| Visual Genome | image objects, relations, attributes manual                    | 42k relations, 3.8M WordNet nodes, 2.3M edges, 2.8M attributes | dog.n.01 | |
| ImageNet  | image objects, crowdsourcing                                    | 14M images, 22k WordNet synsets | |

Observe that the commonsense knowledge is spread over a number of sources with different focus: commonsense knowledge graphs (e.g., ConceptNet), general-domain knowledge graphs (e.g., Wikidata), lexical resources (e.g., WordNet, FrameNet), taxonomies (e.g., Wikidata, WordNet), and visual datasets (e.g., Visual Genome). Therefore, these sources together cover a rich spectrum of knowledge, ranging from everyday knowledge, through event-centric knowledge and taxonomies, to visual knowledge. While the taxonomies have been created manually by experts, most of the commonsense and visual sources have been created by crowdsourcing or curated automatic extraction.\(^1\) Similarly, commonsense and general knowledge graphs tend to be relatively large, with millions of nodes and edges; whereas the taxonomies and the lexical sources are notably smaller. Despite the diverse nature of these sources, we note that many contain mappings to WordNet, as well as a number of other sources. These mappings might be incomplete, e.g., only a small portion of ATOMIC can be mapped to

\(^1\)Commonsense subsets of existing knowledge sources are sometimes also included, e.g., ConceptNet reuses knowledge from Wiktionary and DBpedia.
ConceptNet. Nevertheless, these high-quality mappings provide an opening for consolidation of commonsense knowledge, a goal we pursue in this paper.

2.2 Challenges

Combining these sources in a single graph faces three key challenges:
1. The sources follow different knowledge modeling approaches. One such difference concerns the relation set: there are very few relations in ConceptNet and WordNet, but (tens of) thousands of them in Wikidata and Visual Genome. Consolidation requires a global decision on how to model the relations. The granularity of knowledge is another factor of variance. While regular RDF triples fit some sources (e.g., ConceptNet), representing entire frames (e.g., in FrameNet), event conditions (e.g., in ATOMIC), or compositional image data (e.g., Visual Genome) might benefit from a more open format. An ideal representation would support the entire granularity spectrum.

2. As a number of these sources have been created to support natural language applications, they often contain imprecise descriptions. Natural language phrases are often the main node types in the provided knowledge sources, which provides the benefit of easier access for natural language algorithms, but it introduces ambiguity which might be undesired from a formal semantics perspective. An ideal representation would consolidate various phrasings that share a concept or a referent, while still allowing easy and efficient access to these concepts based on their natural language labels or aliases.

3. Although these sources contain links to existing ones, we observe sparse overlap. As these external links are typically to WordNet, and vary in terms of their version (3.0 or 3.1) or target (lemma or synset), the sources are still disjoint and establishing (identity) connections is difficult. Bridging these gaps, through optimally leveraging existing links, or extending them with additional ones automatically, is a modeling and integration challenge.

Previous efforts that combine commonsense resources exist. A unidirectional manual mapping from VerbNet classes to WordNet and FrameNet is provided by the Unified Verb Index [31]. The Predicate Matrix [6] has a full automatic mapping between lexical resources, including FrameNet, WordNet, and VerbNet. PreMOn [5] formalizes these in RDF. In [19], the authors produce partial mappings between WordNet and Wikipedia/DBpedia. BabelNet [21] integrates several commonsense sources, but extending BabelNet is prevented by its license. Recent systems integrate parts of these sources in an ad-hoc manner to reason on a downstream task, e.g., [35] combine edges from Visual Genome, WordNet, and ConceptNet in a neural network that produces a scene graph from an image.

3 The Common Sense Knowledge Graph

3.1 Principles

To address the aforementioned challenges, we devise principles and a respective representation format that are driven by pragmatic goals of simplicity, modularity, and utility. Specifically, it should be simple to integrate the graph represented
in this format and its arbitrary subsets in reasoning systems, compute (graph and word) embeddings, run off-the-shelf link prediction tools, and integrate it in existing KG-driven downstream reasoning tools [17,15]. We propose that the construction of a unified CSKG should follow five principles:

**P1. Embrace heterogeneity of nodes** While building CSKG, one should preserve the natural node diversity inherent to the variety of sources considered. This entails blurring the distinction between objects (such as those in Visual Genome or Wikidata), classes (such as those in WordNet or ConceptNet), words (in Roget), actions (in ATOMIC or ConceptNet), frames (in FrameNet), and states (as in ATOMIC). It also allows formal nodes, describing unique objects, to co-exist with fuzzy nodes describing ambiguous lexical expressions.

**P2. Reuse edge types across resources** To support reasoning algorithms like KagNet [15], the set of edge types should be kept to a minimum and reused across resources wherever possible. For instance, the ConceptNet edge type /r/HasProperty could be reused to relate a Visual Genome object (e.g., ‘piano’) to its attributes (e.g., ‘black’).

**P3. Leverage external links** The individual graphs are mostly disjoint according to their formal knowledge. However, high-quality links may exist or may be easily inferred, in order to connect these graphs and enable path finding. For instance, while ConceptNet and Visual Genome do not have direct connections, they both make reference to WordNet synsets. Investing an effort in aligning these WordNet synsets would produce a number of very valuable connections between the two knowledge sources.

**P4. Generate high-quality probabilistic links** Experimenting with inclusion of additional probabilistic links would be beneficial, as it would combat sparsity and help path finding algorithms reason over CSKG. These could be inferred with off-the-shelf link prediction algorithms, or with specialized algorithms (see section 3.3 for an example).

**P5. Enable access to labels** The text typically associated with KG nodes, like labels or aliases, provides application-friendly and human-readable access to the CSKG. It can also help us unify descriptions of the same/similar concept across sources. We need to ensure that the graph format supports easy and efficient natural language queries over this text.

### 3.2 Representation

We model CSKG as a **hyper-relational graph**, describing edges in a tabular KGTK [9] format. We opted for this representation rather than the traditional RDF/OWL2 because it allows us to fulfill our goals (of simplicity and utility) and follow our principles more directly, without compromising on the format. For instance, natural language access (principle P5) to RDF/OWL2 nodes requires graph traversal over its rdfs:label relations. Including both reliable and probabilistic nodes (P3 and P4) would require a mechanism to easily indicate edge weights, which in RDF/OWL2 entails inclusion of blank nodes, and a number of additional edges. Moreover, the simplicity of our tabular format allows us to
use standard off-the-shelf functionalities and mature tooling, like the pandas\(^2\) and graph-tool\(^3\) libraries in Python, or graph embedding tools like [14]. These tools have been conveniently wrapped by the KGTK [9] toolkit, which we directly exploit for our analysis and experiments in section 4.

The edges in CSKG are described by ten columns. Following KGTK’s specification, the primary information about an edge consists of its id, node1, relation, and node2. Next, we include four “lifted” edge columns, which leverage the KGTK functionality of representing triples about the primary elements, such as node1;label or relation;label (label of node1 and of relation). Each edge is completed by two “secondary edge” (qualifier) columns, namely: source, which specifies the source(s) of the edge (e.g., “CN” for ConceptNet), and sentence, which contains the natural language lexicalization of a triple, if given by the original source. As each edge has a unique identifier, we can create an auxiliary KGTK file to describe other, more sparse knowledge about some edges, such as its weight. Further specification can be found at: http://shorturl.at/qRR15.

3.3 Consolidation

Currently, CSKG integrates seven sources: a commonsense knowledge graph ConceptNet, a visual commonsense source Visual Genome, a procedural source ATOMIC, a general-domain source Wikidata, and three lexical sources, WordNet, Roget, and FrameNet. Here, we present our design decisions per source, the mappings that facilitate their integration, and further refinements on CSKG.

3.3.1 Individual sources

**ConceptNet** We keep the original data and representation of ConceptNet 5.7, the latest public dump of this graph. We include all 47 relations, including those which are expected to be deprecated in future versions in ConceptNet.\(^4\)

**ATOMIC** We include the entire ATOMIC knowledge graph, preserving the original nodes and nine relations. We do not transform the relations, as the M-to-N best-effort mapping to ConceptNet relations devised in the original paper [24] would result in information loss with respect to the original ATOMIC triples (e.g., at:xAttr and at:oAttr both map to /r/HasProperty, losing the distinction between ‘personX’ and ‘others’ made by ATOMIC). To enhance lexical matching with other sources in CSKG, we add normalized labels of the events and their attributes: converting them to lowercase, removing references to ‘Person\(^*\)’, and excluding ‘none’ values. With this procedure, a template node in ATOMIC, with an initial label “personX accepts personY’s invitation” gets an additional one: “accepts invitation”.

**FrameNet** We import four node types from the FrameNet ontology: frames, frame elements (FEs), lexical units (LUs), and semantic types (STs). We reuse 5 categories of FrameNet edges: frame-frame (13 edge types), frame-FE (1 edge

---

\(^2\)https://pandas.pydata.org/

\(^3\)https://graph-tool.skewed.de/

\(^4\)https://github.com/commonsense/conceptnet5/wiki/Relations
type), frame-LU (1 edge type), FE-ST (1 edge type), and ST-ST (3 edge types). Following principle P2 on edge type reuse, we map these 19 edge types to 9 relations in ConceptNet, e.g., is causative of is converted to /r/Causes.

**Roget** We include all synonyms and antonyms between words in Roget, by reusing the ConceptNet relations /r/Synonym and /r/Antonym (principle P2).

**Visual Genome** does not have its data formatted in a Semantic Web compliant format. Yet, it has been richly annotated, containing an average of 50 descriptions per image, delineating objects from their attributes and relationships, and disambiguating all objects, attributes, and relationships to WordNet synsets. For the objects in Visual Genome, we follow the approach proposed by WebChild and represent them as WordNet synsets. As a result, ‘red shoe’ and ‘shoe’ would both be labels corresponding to the node wn:shoe.n.01. We model all relationships as proximity relations, by reusing the /r/LocatedNear edge type from ConceptNet. The representation of the attributes is conditioned upon the part-of-speech of their WordNet synset. For verb objects, we use ConceptNet’s /r/CapableOf relation (e.g., this would indicate that a person is capable of driving). For adjectives, we introduce a new relation mw:MayHaveProperty, as an uncertain variant of ConceptNet’s /r/HasProperty relation (e.g., indicating that a car may have property blue or fast). Due to the variance in the knowledge expressed with noun attributes, we currently do not include them in CSKG.

**Wikidata** We include the Wikidata-CS subset of Wikidata, extracted in our prior work [11]. This subset of Wikidata has been extracted by following principles of common concepts and generic relations, resulting in 101k statements. Their 45 Wikidata relations were manually mapped to 15 ConceptNet relations.

**WordNet** We include four common relations from WordNet v3.0 by mapping them to three ConceptNet relations: hypernymy (using /r/IsA), part and member holonymy (through /r/PartOf), and substance meronymy (with /r/MadeOf).

### 3.3.2 Mappings

Following our principles, we perform initial node resolution by applying existing identity mappings (P3) and generating probabilistic mappings automatically (P4). We introduce a dedicated relation, mw:SameAs, to indicate identity between two nodes.

**WordNet-WordNet** The WordNet v3.1 identifiers in ConceptNet and the WordNet v3.0 synsets from Visual Genome are aligned by leveraging ILL: the WordNet InterLingual Index. The generated 117,097 mappings are expressed through our identity relation, mw:SameAs.

**WordNet-Wikidata** We compute probabilistic links between WordNet synsets and Wikidata taxonomy nodes. Our approach consists of three components: a Candidate Retrieving Module (CRM), a Similarity Calculating Module (SCM), and a Mapping Module (MM). CRM retrieves candidate nodes from a customized ElasticSearch index of Wikidata. Concretely, it matches a synset word in any text field (including labels, aliases, and descriptions), and it ranks the candidates with a version of the default TF-IDF-based algorithm which adapts the

---

5 https://github.com/globalwordnet/ili
score proportionally to the number of incoming links. The top-ranked \( n = 50 \) candidates are retained. Then, SCM computes sentence embeddings of the descriptions of the WordNet synset and each of the Wikidata candidates by using a pre-trained XLNet model [34]. The similarity between a synset and a Wikidata node is computed as a cosine similarity between their corresponding embeddings. MM creates a \texttt{mw:SameAs} edge between a WordNet synset and the Wikidata candidate with highest similarity. This similarity is represented as weight of the mapping edge. The accuracy of each mapping has been validated by one student. In total, 17 students took part in this validation. Out of all edges produced by the algorithm (112,012), the manual validation marked 57,145 as correct. We keep these in CSKG and discard the rest.

**FrameNet-ConceptNet** We connect the FrameNet nodes to ConceptNet in two ways. Its lexical units are mapped to corresponding ConceptNet nodes through the Predicate Matrix (cf. section 2), producing 3,016 \texttt{mw:SameAs} edges. Then, we use \( \approx 200k \) hand-annotated sentences from the FrameNet corpus, each annotated with its target frame, a set of FEs, and the words associated with each FE. We treat these words as lexical units of the corresponding FE. We ground these sets of words to ConceptNet with the rule-based method of [15], thus generating 45,659 \texttt{fn:HasLexicalUnit} edges.

**Lexical matching** We establish 74,259 \texttt{mw:SameAs} links between nodes in ATOMIC, ConceptNet, and Roget by exact lexical match of their labels. We restrict this matching to lexical nodes (e.g., /c/en/cat and not /c/en/cat/n/wn/animal).

### 3.3.3 Refinement
We consolidate the seven sources and their interlinks as follows. After transforming them to the representation described in the past two sections, we concatenate them in a single graph. We deduplicate this graph and append all mappings, resulting in CSKG*. Finally, we apply the mappings to merge identical nodes (connected with \texttt{mw:SameAs}) and perform a final deduplication of the edges, resulting in our consolidated CSKG graph. The entire procedure of importing the individual sources and consolidating them into CSKG is implemented with KGTK operations [9], and can be found on our GitHub.  

### 3.4 Embeddings
Embeddings provide a convenient entry point to knowledge graphs and enable reasoning on both intrinsic and downstream tasks. For instance, many reasoning applications (cf. [17,15]) of ConceptNet rely of their official NumberBatch embeddings [27].

Motivated by these observations, we aspire to produce high-quality embeddings of the CSKG graph. We experiment with two families of embedding algorithms. On the one hand, we produce variants of popular graph embeddings:

---

6. Documentation: shorturl.at/jpBY3

7. During this step, we manually fixed approximately 250 nodes which contained spelling and tokenization errors in the Predicate Matrix.

8. https://github.com/usc-isi-i2/cskg/blob/master/consolidation/create_cskg.sh
Table 2. CSKG statistics. Abbreviations: CN=ConceptNet, VG=Visual Genome, WN=WordNet, RG=Roget, WD=Wikidata, FN=FrameNet, AT=ATOMIC. Relation numbers in brackets are before consolidating to ConceptNet.

| AT | CN  | FN   | RG   | VG   | WD   | WN   | CSKG* | CSKG |
|----|-----|------|------|------|------|------|-------|------|
| #nodes | 304,909 | 1,787,373 | 15,652 | 71,804 | 91,294 | 71,243 | 2,414,813 | 2,160,968 |
| #edges | 732,723 | 3,423,004 | 29,873 | 1,403,955 | 2,587,623 | 111,276 | 101,771 | 6,349,731 | 6,001,531 |
| #relations | 9 | 47 | 9 (23) | 2 | 3 (42k) | 3 | 15 (45) | 59 | 58 |
| avg degree | 4.81 | 3.83 | 3.82 | 39.1 | 459.45 | 2.44 | 2.86 | 5.26 | 5.55 |
| std degree | 0.07 | 0.02 | 0.13 | 0.34 | 35.81 | 0.02 | 0.05 | 0.02 | 0.03 |

Table 3. Nodes with highest centrality score according to PageRank and HITS. Node labels indicated in bold.

| PageRank | HITS hubs | HITS authorities |
|----------|-----------|------------------|
| /c/en/chromatic/a/wn | /c/en/red | /c/en/blue |
| /c/en/organic_compound | /c/en/yellow | /c/en/red |
| /c/en/chemical_compound/n | /c/en/green | /c/en/silver |
| /c/en/change/n/wn/artifact | /c/en/silver | /c/en/green |
| /c/en/natural_science/n/wn/cognition | /c/en/blue | /c/en/gold |

TransE [4], DistMult [33], ComplEx [30], and RESCAL [23]. On the other hand, we produce various text (Transformer-based) embeddings based on BERT-large [8]. For BERT, we first create a sentence for each node, based on a template that encompasses its neighborhood, which is then encoded with BERT’s sentence transformer model. All embeddings are computed with the KGTK operations graph-embeddings and text-embeddings. We analyze them in section 4.2.

The CSKG embeddings are publicly available at http://shorturl.at/pAGX8.

4 Analysis

4.1 Statistics

Basic statistics are shown in Table 2. In total, our mappings produce 251,517 mw:SameAs links and 45,659 mw:HasInstance links. After refinement, i.e., removal of the duplicates and merging of the identical nodes, CSKG consists of 2.2 million nodes and 6 million edges. In terms of edges, its largest subgraph is ConceptNet (3.4 million), whereas ATOMIC comes second with 733 thousand edges. These two graphs also contribute the largest number of nodes to CSKG. The three most common relations in CSKG are: /r/RelatedTo (1.7 million), /r/Synonym (1.2 million), and /r/Antonym (401 thousand edges).

Connectivity and centrality The mean degree of CSKG grows by 5.5% (from 5.26 to 5.55) after merging identical nodes. Compared to ConceptNet, its degree is 45% higher, due to its increased number of edges while keeping...
Table 4. Top-5 most similar nodes for /c/en/turtle/n/wn/animal (E1) and /c/en/happy (E2) according to TransE and BERT.

| TransE                      | BERT                      |
|-----------------------------|---------------------------|
| E1 /c/en/chelonian/n/wn/animal | /c/en/glyptemys/n         |
| /c/en/mud_turtle/n/wn/animal   | /c/en/pelocomastes/n      |
| /c/en/cooter/n/wn/animal      | /c/en/staurotypus/n       |
| /c/en/common_snapping_turtle/n/wn/animal | /c/en/parahydraspis/n     |
| /c/en/sea_turtle/n/wn/animal  | /c/en/trachemys/n         |
| E2 /c/en/excited             | /c/en/bring_happiness     |
| /c/en/satisfied              | /c/en/new_happiness       |
| /c/en/smile_mood             | at:like_a_party_is_a_good_way_to... |
| /c/en/pleased                | /c/en/encouraging_person’s_talent |
| /c/en/joyful                 | at:happy_that_they_went_to_the_party |

the number of nodes nearly constant. The best connected subgraphs are Visual Genome and Roget. CSKG’s high connectivity is owed largely to these two sources and our mappings, as the other five sources have degrees below that of CSKG. The abnormally large node degrees and variance of Visual Genome are due to its annotation guidelines that dictate all concept-to-concept information to be annotated, and our modeling choice to represent its nodes through their synsets. We report that the in-degree and out-degree distributions of CSKG have Zipfian shapes, a notable difference being that the maximal in degree is nearly double compared to its maximal out degree (11k vs 6.4k). To understand better the central nodes in CSKG, we compute PageRank and HITS metrics. The top-5 results are shown in Table 3. We observe that the node with highest PageRank has label “chromatic”, while all dominant HITS hubs and authorities are colors, revealing that knowledge on colors of real-world object is common in CSKG. PageRank also reveals that knowledge on natural and chemical processes is well-represented in CSKG. Finally, we note that the top-centrality nodes are generally described by multiple subgraphs, e.g., /c/en/natural_science/n/wn/cognition is found in ConceptNet and WordNet, whereas the color nodes (e.g., /c/en/red) are shared between Roget and ConceptNet.

4.2 Analysis of the CSKG embeddings

We randomly sample 5,000 nodes from CSKG and visualize their embeddings computed with an algorithm from each family: TransE and BERT. The results are shown in Figure 1. We observe that graph embeddings group nodes from the same source together. This is because graph embeddings tend to focus on the graph structure, and because most links in CSKG are still within sources. We observe that the sources are more intertwined in the case of the BERT embeddings, because of the emphasis on lexical over structural similarity. Moreover, in both plots Roget is dispersed around the ConceptNet nodes, which is likely due to its broad coverage of concepts, that maps both structurally and lexically
Fig. 1. UMAP visualization of 5,000 randomly sampled nodes from CSKG, represented by TransE (left) and BERT (right) embeddings. Colors signify node sources.

to ConceptNet. At the same time, while ATOMIC overlaps with a subset of ConceptNet, the two sources mostly cover different areas of the space.

Table 4 shows the top-5 most similar neighbors for /c/en/turtle/n/wn/animal and /c/en/happy according to TransE and BERT. We note that while graph embeddings favor nodes that are structurally similar (e.g., /c/en/turtle/n/wn/animal and /c/en/chelonian/n/wn/animal are both animals in WordNet), text embeddings give much higher importance to lexical similarity of nodes or their neighbors, even when the nodes are disconnected in CSKG (e.g., /c/en/happy and at:happy that they went to the party). These results are expected considering the approach behind each algorithm.

Word association with embeddings To quantify the utility of different embeddings, we evaluate them on the USF-FAN [22] benchmark, which contains crowdsourced common sense associations for 5,019 “stimulus” concepts in English. For instance, the associations provided for day are: night, light, sun, time, week, and break. The associations are ordered descendingly based on their frequency. With each algorithm, we produce a top-K most similar neighbors list based on the embedding of the stimulus concept. Here, K is the number of associations for a concept, which varies across stimuli. If CSKG has multiple nodes for the stimulus label, we average their embeddings. For the graph embeddings, we use logistic loss function, using a dot comparator, a learning rate of 0.1, and dimension 100. The BERT text embeddings have dimension 1024, which is the native dimension of this language model. As the text embedding models often favor surface form similarity (e.g., associations like daily for day), we devise variants of this method that excludes associations with Levenshtein similarity higher than a threshold t.

We evaluate by comparing the embedding-based list to the benchmark one, through customary ranking metrics, like Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG). Our investigations show that TransE is the best-performing algorithm overall, with MAP of 0.207 and NDCG of 0.530. The optimal BERT variant uses threshold of t = 0.9, scoring with MAP.
Table 5. Number of triples retrieved with ConceptNet and CSKG on different datasets. #Q=number of questions.

|       | #Q | ConceptNet | CSKG | #Q | ConceptNet | CSKG |
|-------|----|------------|------|----|------------|------|
| CSQA  | 9,741 | 78,729 | 125,552 | 1,221 | 9,758 | 15,662 |
| SIQA  | 33,410 | 126,596 | 266,937 | 1,954 | 7,850 | 16,149 |
| PIQA  | 16,113 | 18,549 | 59,684 | 1,838 | 2,170 | 6,840 |
| aNLI  | 169,654 | 257,163 | 638,841 | 1,532 | 5,603 | 13,582 |

of 0.209 and NDCG of 0.268. In the future, we aim to investigate embedding techniques that can seamlessly integrate structural and content information.

5 Applications

As the creation of CSKG is largely driven by downstream reasoning needs, we now investigate its relevance for commonsense question answering: 1) we measure its ability to contribute novel evidence to support reasoning, and 2) we measure its role in pre-training language models for zero-shot downstream reasoning.

5.1 Retrieving evidence from CSKG

We measure the relevance of CSKG for commonsense question answering tasks, by comparing the number of retrieved triples that connect keywords in the question and in the answers. For this purpose, we adapt the lexical grounding in HyKAS [17] to retrieve triples from CSKG instead of its default knowledge source, ConceptNet. We expect that CSKG can provide much more evidence than ConceptNet, both in terms of number of triples and their diversity. We experiment with four commonsense datasets: CommonSense QA (CSQA) [28], Social IQA (SIQA) [25], Physical IQA (PIQA) [3], and abductive NLI (aNLI) [2].

As shown in Table 5, CSKG significantly increases the number of evidence triples that connect terms in questions with terms in answers, in comparison to ConceptNet. We note that the increase is on average 2-3 times, the expected exception being CSQA, which was inferred from ConceptNet.

We inspect a sample of questions to gain insight into whether the additional triples are relevant and could benefit reasoning. For instance, let us consider the CSQA question “Bob the lizard lives in a warm place with lots of water. Where does he probably live?”, whose correct answer is “tropical rainforest”. In addition to the ConceptNet triple /c/en/lizard /c/en/AtLocation /c/en/tropical_rainforest, CSKG provides two additional triples, stating that tropical is an instance of place and that water may have property tropical. The first additional edge stems from our mappings from FrameNet to ConceptNet, whereas the second comes from Visual Genome. We note that, while CSKG increases the coverage with respect to available commonsense knowledge,
Table 6. Zero-shot evaluation results with different combinations of models and knowledge sources, across five commonsense tasks, as reported in [18]. CWWV is a QA set that combines ConceptNet, Wikidata, WordNet, and Visual Genome. CSKG is a union of ATOMIC and CWWV. We run our experiments three times with different seeds and report average accuracy with 95% confidence interval.

| Model      | KG      | aNLI    | CSQA    | PIQA    | SIQA    | WG       |
|------------|---------|---------|---------|---------|---------|----------|
| GPT2-L     | ATOMIC  | 59.2(±0.3) | 48.0(±0.9) | 67.5(±0.7) | 53.5(±0.4) | 54.7(±0.6) |
| GPT2-L     | CWWV    | 58.3(±0.4) | 46.2(±1.0) | 68.6(±0.7) | 48.0(±0.7) | 52.8(±0.9) |
| GPT2-L     | CSKG    | 59.0(±0.5) | 48.6(±1.0) | 68.6(±0.9) | 53.3(±0.5) | 54.1(±0.5) |
| RoBERTa-L  | ATOMIC  | 70.8(±1.2) | 64.2(±0.7) | 72.1(±0.5) | 63.1(±1.5) | 59.6(±0.3) |
| RoBERTa-L  | CWWV    | 70.0(±0.3) | 67.9(±0.8) | 72.0(±0.7) | 54.8(±1.2) | 59.4(±0.5) |
| RoBERTa-L  | CSKG    | 70.5(±0.2) | 67.4(±0.8) | 72.4(±0.4) | 63.2(±0.7) | 60.9(±0.8) |
| **Human**  |         | 91.4     | 88.9    | 94.9    | 86.9    | 94.1     |

it is also incomplete: in the above example, useful information such as warm temperatures being typical for tropical rainforests is still absent.

5.2 Fine-tuning language models with CSKG

We have studied the role of various subsets of CSKG for downstream QA reasoning extensively in [18]. Here, CSKG or its subsets were transformed into artificial commonsense question answering tasks. These tasks were then used instead of training data to pre-train language models, like RoBERTa and GPT-2. Such a CSKG-based per-trained language model was then ‘frozen’ and evaluated in a zero-shot manner across a wide variety of commonsense tasks, ranging from question answering through pronoun resolution and natural language inference.

We select key results from these experiments in Table 6. The results demonstrate that no single knowledge source suffices for all benchmarks and that using CSKG is overall beneficial compared to using its subgraphs. The kind of knowledge that benefits each task is ultimately conditioned on the alignment between this knowledge and the targeted task, indicating that subsequent work should further investigate how to dynamically align knowledge with the task at hand.

6 Discussion

Our analysis in section 4 revealed that the connectivity in CSKG is higher than merely concatenation of the individual sources, due to our mappings across sources and the merge of identical nodes. Its KGTK format allowed us to seamlessly compute and evaluate a series of embeddings, observing that TransE and BERT with additional filtering are the two best-performing and complementary algorithms. Furthermore, the novel evidence brought by CSKG on downstream QA tasks (section 5) is a signal that can be exploited by reasoning systems to
enhance their performance and robustness, as shown in [18]. Yet, the quest to a rich, high-coverage CSKG is far from completed:

**Node resolution** As large part of CSKG consists of lexical nodes, it suffers from the standard challenges of linguistic ambiguity and variance. For instance, there are 18 nodes in CSKG that have the label ‘scene’, which includes WordNet or OpenCyc synsets, Wikidata Qnodes, frame elements, and a lexical node. Variance is another challenge, as `/c/en/caffine`, `/c/en/caffeine`, and `/c/en/the active ingredient caffeine` are all separate nodes in ConceptNet (and in CSKG). We are currently investigating techniques for node resolution over CSKG to address these challenges.

**Semantic enrichment** We have normalized the edge types across sources to a single, ConceptNet-centric, set of 58 relations. In [10], we classify all CSKG’s relations into 13 dimensions, enabling us to consolidate the edge types further. At the same time, some of these relations hide fine-grained distinctions, for example, WebChild [29] defines 19 specific property relations, including temperature, shape, and color, all of which correspond to ConceptNet’s `/r/HasProperty`. A novel future direction is to produce hierarchy for each of the relations, and refine existing triples by using a more specific relation (e.g., use the predicate ‘temperature’ instead of ‘property’ when the object of the triple is ‘cold’).

## 7 Conclusions and Future Work

While current commonsense knowledge sources contain complementary knowledge that would be beneficial as a whole for downstream tasks, such usage is prevented by different modeling approaches, foci, and sparsity of available mappings. Optimizing for simplicity, modularity, and utility, we proposed a hyper-relational graph representation that describes many nodes with a few edge types, maximizes the high-quality links across subgraphs, and enables natural language access. We applied this representation approach to consolidate a commonsense knowledge graph (CSKG) from seven very diverse and disjoint sources: a text-based commonsense knowledge graph ConceptNet, a general-purpose taxonomy Wikidata, an image description dataset Visual Genome, a procedural knowledge source ATOMIC, and three lexical sources: WordNet, Roget, and FrameNet. CSKG describes 2.2 million nodes with 6 million statements. Our analysis showed that CSKG is a well-connected graph and more than ‘a simple sum of its parts’. We also release a series of graph and text embeddings of the CSKG nodes, to facilitate future usage of the graph. Our analysis showed that graph and text embeddings of CSKG have complementary notions of similarity, as the former focus on structural patterns, while the latter on lexical features of the node’s label and of its neighborhood. Applying CSKG on downstream commonsense reasoning tasks, like QA, showed an increased recall as well as an advantage when training a language model to reason across datasets in a zero-shot fashion. Key standing challenges for CSKG include semantic consolidation of its nodes and refinement of its property hierarchy. CSKG and its embeddings are publicly available. Note-
books for analyzing these resources can be found on our public GitHub page: https://github.com/usc-isi-i2/cskg/tree/master/ESWC2021.

References

1. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley framenet project. In: Proceedings of the 17th international conference on Computational linguistics-Volume 1. pp. 86–90. Association for Computational Linguistics (1998)
2. Bhagavatula, C., Bras, R.L., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., Yih, S.W.t., Choi, Y.: Abductive commonsense reasoning. arXiv preprint arXiv:1908.05739 (2019)
3. Bisk, Y., Zellers, R., Bras, R.L., Gao, J., Choi, Y.: Piqa: Reasoning about physical commonsense in natural language. arXiv preprint arXiv:1911.11641 (2019)
4. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. Advances in neural information processing systems 26, 2787–2795 (2013)
5. Corcoglioniti, F., Rospocher, M., Aprosio, A.P., Tonelli, S.: Premon: a lemon extension for exposing predicate models as linked data. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16). pp. 877–884 (2016)
6. De Lacalle, M.L., Laparra, E., Aldabe, I., Rigau, G.: Predicate matrix: automatically extending the semantic interoperability between predicate resources. Language Resources and Evaluation 50(2), 263–289 (2016)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
8. 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)
9. 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)
10. Ilievski, F., Oltramari, A., McGuinness, D., Szekely, P., Ma, K., Zhang, B.: Dimensions of commonsense knowledge. Knowledge-Based Systems (in prep)
11. Ilievski, F., Szekely, P., Schwabe, D.: Commonsense knowledge in wikidata. arXiv preprint arXiv:2008.08114 (2020)
12. Kipfer, B.: Roget’s 21st century thesaurus in dictionary form (éd. 3). new york: The philip lief group (2005)
13. 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)
14. Lerer, A., Wu, L., Shen, J., Lacroix, T., Wehrstedt, L., Bose, A., Peysakhovich, A.: Pytorch-biggraph: A large-scale graph embedding system. arXiv preprint arXiv:1903.12287 (2019)
15. Lin, B.Y., Chen, X., Chen, J., Ren, X.: Kagnet: Knowledge-aware graph networks for commonsense reasoning. arXiv preprint arXiv:1909.02151 (2019)
16. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (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. Ma, K., Ilievski, F., Francis, J., Bisk, Y., Nyberg, E., Oltramari, A.: Knowledge-driven self-supervision for zero-shot commonsense question answering. arXiv preprint arXiv:2011.03863 (2020)
19. McCrae, J.P.: Mapping wordnet instances to wikipedia. In: Proceedings of the 9th Global WordNet Conference (GWC 2018). pp. 62–69 (2018)
20. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM 38(11), 39–41 (1995)
21. Navigli, R., Ponzetto, S.P.: Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193, 217–250 (2012)
22. Nelson, D.L., McEvoy, C.L., Schreiber, T.A.: The university of south florida free association, rhyme, and word fragment norms. Behavior Research Methods, Instruments, & Computers 36(3), 402–407 (2004)
23. Nickel, M., Tresp, V., Kriegel, H.P.: A three-way model for collective learning on multi-relational data. In: Icml. vol. 11, pp. 809–816 (2011)
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. Sap, M., Rashkin, H., Chen, D., LeBras, R., Choi, Y.: Socialiqa: Commonsense reasoning about social interactions. arXiv preprint arXiv:1904.09728 (2019)
26. Schuler, K.K.: Verbnet: A broad-coverage, comprehensive verb lexicon (2005)
27. 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)
28. Talmor, A., Herzig, J., Lourie, N., Berant, J.: Commonsenseqa: A question answering challenge targeting commonsense knowledge. arXiv preprint arXiv:1811.00937 (2018)
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. Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G.: Complex embeddings for simple link prediction. International Conference on Machine Learning (ICML) (2016)
31. Trumbo, D.: Increasing the usability of research lexica. Ph.D. thesis, University of Colorado at Boulder (2006)
32. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM 57(10), 78–85 (2014)
33. Yang, B., Yih, W.t., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575 (2014)
34. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems. pp. 5754–5764 (2019)
35. Zareian, A., Karaman, S., Chang, S.F.: Bridging knowledge graphs to generate scene graphs. arXiv preprint arXiv:2001.02314 (2020)
36. Zellers, R., Bisk, Y., Schwartz, R., Choi, Y.: Swag: A large-scale adversarial dataset for grounded commonsense inference. arXiv preprint arXiv:1808.05326 (2018)