Commonsense Knowledge in Word Associations and ConceptNet

Chunhua Liu **Trevor Cohn** **Lea Frermann**
School of Computing and Information Systems
The University of Melbourne
chunhua@student.unimelb.edu.au
{tcohn,lfrermann}@unimelb.edu.au

Abstract

Humans use countless basic, shared facts about the world to efficiently navigate in their environment. This *commonsense knowledge* is rarely communicated explicitly, however, understanding how commonsense knowledge is represented in different paradigms is important for both deeper understanding of human cognition and for augmenting automatic reasoning systems. This paper presents an in-depth comparison of two large-scale resources of general knowledge: ConceptNet, an engineered relational database, and SWOW a knowledge graph derived from crowd-sourced word associations. We examine the structure, overlap and differences between the two graphs, as well as the extent to which they encode situational commonsense knowledge. We finally show empirically that both resources improve downstream task performance on commonsense reasoning benchmarks over text-only baselines, suggesting that large-scale word association data, which have been obtained for several languages through crowd-sourcing, can be a valuable complement to curated knowledge graphs.[1]

1 Introduction

Humans understand and navigate everyday situations with great efficiency using a wealth of shared, basic facts about their social and physical environment – a resource often called commonsense knowledge (Liu and Singh, 2004). Advances in artificial intelligence in general, and natural language processing in particular, have led to a surge of interest in its nature: what constitutes commonsense knowledge? And despite this knowledge being rarely explicitly stated in text (Gordon and Durme, 2013), how can we equip machines with commonsense to enable more general inference (Davis and Marcus, 2015)? Recently, large language models (Devlin et al., 2019; Radford et al., 2019) pretrained on massive text corpora achieved promising results on commonsense reasoning benchmarks with fine-tuning, however, the lack of interpretability remains a problem. Augmenting them with commonsense knowledge can provide complementary knowledge (Ilievski et al., 2021a; Safavi and Koutra, 2021) and thus make models more robust and explainable (Lin et al., 2019).

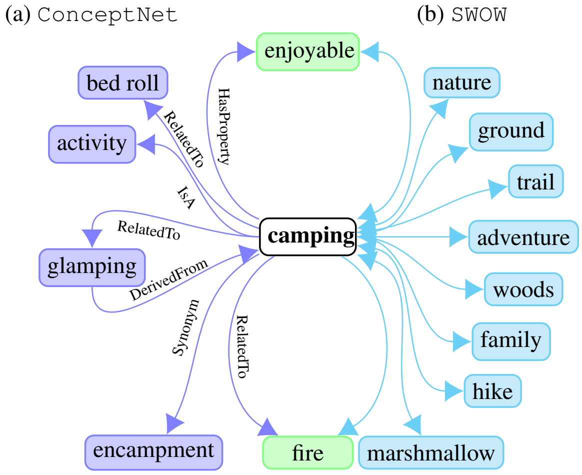


Figure 1: Sub-graphs centered around '**camping**' from (a) ConceptNet and (b) SWOW. Nodes in green are common to both KGs. Nodes on the left/blue (right/cyan) are unique to ConceptNet (SWOW).

Attempts to capture general human knowledge include inventories of machine-readable logical rules (Gordon and Hobbs, 2017) and large, curated databases which have been collected with the specific purpose to reflect either domain general (e.g., ConceptNet; Liu and Singh (2004)) or domain-specific (e.g., ATOMIC; Sap et al. (2019)) commonsense knowledge.

Word association norms are a third resource of explicit, basic human knowledge: in a typical study, human participants are presented with a *cue* word (e.g., 'camping') and asked to produce one or

[1] Code available at https://github.com/ChunhuaLiu596/CSWordAssociation

Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL), pages 481–495
November 10–11, 2021. ©2021 Association for Computational Linguistics

more words that spontaneously come to mind (e.g., 'hike', 'nature', or 'trail'). The resulting data sets of word associations have been used to explore the mental representations and connections underlying human language and reasoning (Nelson et al., 2004; Fitzpatrick, 2006), and have been shown to contain information beyond that present in text corpora (De Deyne et al., 2016). Word association studies have been scaled to thousands of cue words, tens of thousands of participants, and several languages, and hence provide a way of collecting diverse and unbiased representations. However, the extent to which they capture *commonsense* knowledge, and their utility for downstream applications have not yet been examined.

Prior work has probed the extent to which word associations contain lexical knowledge (De Deyne et al., 2016; Deyne et al., 2019), has systematically compared the relational coverage and overlap of curated knowledge bases (Ilievski et al., 2020) (not including word associations data), and has investigated how much relational commonsense knowledge is present in language models (A. Rodriguez and Merlo, 2020; Vankrunkelsven et al., 2018; Da and Kasai, 2019). We contribute to this line of research with an in-depth comparison of a dedicated commonsense knowledge base, and resources derived from spontaneous human-generated word associations. We systematically compare the most comprehensive, domain general, curated commonsense knowledge base (ConceptNet) with the largest data set of English word associations (the "Small World of Words"; SWOW; Deyne et al. (2019)). We compare the two resources both in their structure and content, and apply them in downstream reasoning tasks.

Our contribution is important for three reasons. First, comparing explicitly engineered with spontaneously produced knowledge graphs can advance our fundamental understanding of the similarities and differences between the paradigms, and suggest ways to combine them. Second, recent progress in automatic commonsense reasoning largely focused on English and relies heavily on the availability of very large language models. These are, however, infeasible to train for all but a few high-resource languages. Finally, recent work has shown that the competitive performance of large language models on commonsense reasoning tasks is at least partially due to spurious correlations in language rather than genuine reasoning abilities; and that they perform close to random once evaluation controls for such confounds (Elazar et al., 2021). Explicit representations of commonsense knowledge bases are thus have the potential to promote robust and inclusive natural language reasoning models.

In summary, our contributions are:

1. We conduct an in-depth comparison of large-scale curated commonsense knowledge bases and word association networks, distilling a number of systematic differences which can inform future theoretical and empirical work.

2. We analyze how much *commonsense* knowledge ConceptNet and SWOW encode, leveraging a human-created data set covering explicit situational knowledge. Our results suggest that SWOW represents this knowledge more directly.

3. We introduce SWOW as a commonsense resource for NLP applications and show that it achieves comparable results with ConceptNet across three commonsense question answering benchmarks.

2 Background

2.1 Commonsense Knowledge Graphs

Collecting and curating human commonsense knowledge has a rich history in artificial intelligence and related fields, motivated by the grand challenge of equipping AI systems with commonsense knowledge and reasoning ability (Davis and Marcus, 2015; Lake et al., 2017). Previous efforts have been put on collecting different aspects of commonsense knowledge, resulting in a variety of resources ranging from systems of logical rules (Lenat and Guha, 1993), over knowledge graphs (Liu and Singh, 2004), all the way to embedded representations (Malaviya et al., 2020). Here, we focus on graph representations. Domain-specific examples include ATOMIC (Sap et al., 2019) which focuses on social interactions in events, SenticNet (Cambria et al., 2020) which encodes sentiment-related commonsense knowledge. Recent work also attempts to consolidate knowledge from multiple sources in order to improve knowledge coverage and utility (Ilievski et al., 2021b; Hwang et al., 2021). The largest domain-general commonsense knowledge graph is ConceptNet, which we will use in this study and describe in detail below.

KG	#Triples	#Nodes	#Relations	Density	Degree	H_N
ConceptNet	3,009,636	1,080,759	47	3.00×10^{-6}	2.78	23.28
SWOW	1,593,564	124,626	2	1.03×10^{-4}	12.78	18.07

Table 1: Statistics of ConceptNet and SWOW considered as directed graphs. Density is the graph density, Degree indicates the average node degree. H_N indicates the node entropy.

ConceptNet All studies in this paper are based on the most recent ConceptNet v5.6 (Speer et al., 2017). ConceptNet is a directed graph comprising over 3M nodes (aka concepts). Related concepts are connected with directed edges, which are labelled with one of 47 generic relation types. Figure 1a shows a small subgraph, centered around the concept 'camping'. Nodes are represented as free-text descriptions, which leads to a large node inventory and a sparsely connected graph. We filter out nodes that are not English, lowercase all descriptions and remove punctuation. Row 1 in Table 1 shows statistics of the resulting knowledge graph.

2.2 Word Association Networks

Human word associations have a long history in psychology and cognitive linguistics (Deese, 1966; Deyne et al., 2019). Given a *cue* word, one or more spontaneous *responses* are elicited from study participants, shedding light on their mental associations. The resulting data sets, covering many cues and participants, can subsequently be turned into association *networks*, where each node corresponds to either a cue or response. Cues are connected to the responses they elicited through directed edges, which can be weighted (or filtered) by the number of participants who produced a particular response. The primary use-case of word associations has been to gain insights into the mental lexicon. A wealth of studies has shown the utility of word associations for predicting behavioural data including memory, lexical choice and semantic categorization (Nelson et al., 2000; De Deyne et al., 2013; Borge-Holthoefer and Arenas, 2010), however, we are the first to inspect word association networks as a general commonsense knowledge base. Several association data sets have been collected, varying in coverage and target language (Kiss et al. (1973); Nelson et al. (2004); Jung et al. (2010); Simon De Deyne and Storms (2013)). However, in order to consider this data as a general knowledge resource, it should (a) cover a large set of diverse cues, and (b) a large number of responses which are both diverse and reliable. Recently, the *Small World of Words* project massively scaled word association collection through crowd-sourcing (Deyne et al., 2019). In the remainder of this paper, we use their English data set ("SWOW", described below). The *Small World of Words* project has been scaled to 15 languages, suggesting its potential as a knowledge resource for NLP more generally.

Small World of Words SWOW (Deyne et al., 2019) is a word association network derived from a large collection of crowd-sourced cue-response pairs involving more than 90K participants and 12K cues. Given a cue, participants produced up to three responses. The resulting associations have been compiled into the SWOW knowledge graph (Figure 1b shows a small excerpt). SWOW edges are not labelled with relations. In some of our experiments, we adopt a basic set of two relations using the associative directions, namely *forward associations* from a cue to a response, and *mutual associations* for pairs where the reverse is also included in SWOW. We use the official, pre-processed release of SWOW.[2] We remove "NA" responses, lowercase all node descriptions and remove punctuation. Row 2 in Table 1 shows statistics of the resulting graph.

In the remainder of this paper, we conduct three experiments to compare SWOW and ConceptNet from different dimensions, ranging from the intrinsic graph properties (§3), to the coverage of encoded commonsense knowledge (§4), and their utility for downstream commonsense reasoning tasks (§5).

3 Experiment 1: Intrinsic Comparisons

A knowledge graph consists of a set of nodes $\mathcal{N}$ and edges $\mathcal{E}$, comprising triples $\langle e, r, e' \rangle$ to denote a directed edge from head node e to tail node e' labelled with relation $r \in \mathcal{R}$. We denote $\mathcal{E}(e)$ as the incoming and outgoing edge set for node e, $\mathcal{E}(r)$ as the set of edges with relation r, and $|\cdot|$ as the size of a set. Here, we consider the specific

[2] https://smallworldofwords.org/en/project/research

knowledge graphs ConceptNet and SWOW, and begin by comparing the intrinsic properties: their typology and content encoded.

3.1 Knowledge Graph Structure

ConceptNet is a substantially larger graph than SWOW, with about eight times as many nodes and 1.9× as many edges (cf., Table 1). We compare sparsity in terms of (1) *graph density*,

$$\frac{|\mathcal{E}|}{|\mathcal{N}|(|\mathcal{N}|-1)},$$

and (2) *node degree* as the average total of incoming and outgoing edges (Malaviya et al., 2020). Table 1 shows that SWOW has 39× the density and 4× the average node degree of ConceptNet: Even though SWOW is smaller than ConceptNet, it is substantially more densely connected.

Node Distribution. To better understand the distribution of nodes in the KGs, we measure node diversity via the entropy of the node distribution (H_N; Pujara et al. (2017)):

$$H_N = \sum_{e \in N} -P(e) \log P(e),$$

where $P(e) = |\mathcal{E}(e)|/|\mathcal{E}|$ is the fraction of edges incident on the node. Higher H_N indicates a more uniform node distribution where many nodes are connected, whereas lower entropy suggests a skewed distribution, where few nodes are highly connected. Table 1 shows a lower H_N for SWOW, i.e., nodes are less uniformly connected. This is because SWOW is by construction more structured than ConceptNet: its 12K *cue* nodes are densely connected to a much larger number of (sparsely connected) *response* nodes.

Node and Edge Overlap. How large is the overlap between ConceptNet and SWOW? We quantify the overlap of individual nodes in the two KGs, based on exact string match.[3] We find 58% (71K) of the nodes in the smaller SWOW are present in ConceptNet (conversely, 7% of the nodes in ConceptNet are present in SWOW). Over 40% of the concepts in SWOW are not present in ConceptNet, which is perhaps expected given their very distinct methods of construction, but motivates further in-depth comparison (Section 3.2). Moving on to edge overlap, which we measure over undirected head-tail pairs,[4] we find that 6% of edges in ConceptNet are present in SWOW, and 15% of the edges in SWOW are present in ConceptNet. This low overlap demonstrates that human associations indeed elicit connections among words missed in the large database ConceptNet. For the 71K overlapping nodes, we further find that 691K connections exist in ConceptNet, and 1.5M connections in SWOW, covering 95% of all SWOW edges. This again suggests that SWOW is more comprehensive than ConceptNet.

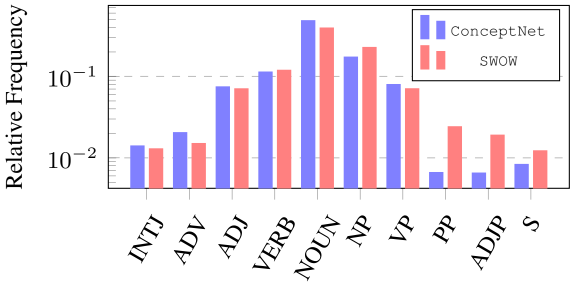


Figure 2: The distribution of syntactic tags on ConceptNet and SWOW for the 10 most frequent tags.

3.2 Knowledge Graph Content

Having established the structural characteristics of ConceptNet and SWOW, we will now focus on their respective encoded knowledge.

3.2.1 Conceptual Content

Nodes in ConceptNet and SWOW express concepts as words or short phrases. We compare: (1) the distributions of the syntactic categories for concepts over the two knowledge bases; (2) the occurrences of concepts in two KGs in large corpus.

We use a constituency parser to predict the syntactic phrase or part of speech (POS) tag for a concept string.[5] The relative prevalence of the 10 most frequent syntactic types is shown in Figure 2. While the overall distribution is similar in both KGs, two patterns emerge. First, even though both KGs are dominated by nominal nodes, SWOW's dis-

[3] String matching is arguably a simplistic way of matching concepts across two KGs. Advanced methods of concept resolution could leverage embedding methods. We leave this interesting direction for future work.

[4] Edge comparison ignores direction as many relations can naturally be inverted, e.g., 'part of' and 'has part'. Consequently linking concepts in either direction is considered to be correct.

[5] We use the parser of Kitaev and Klein (2018) as implemented in Spacy.

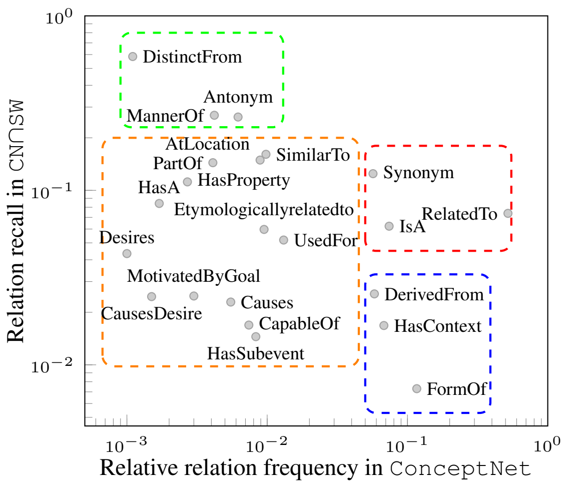


Figure 3: The correlation between the relative frequency of `ConceptNet` relations and their recall in the overlap subgraph, `CN∩SW`.

tribution over POS types is less skewed, suggesting concepts are more diverse. Secondly, the proportion of phrasal concepts compared to single-word concepts tends to be higher in `SWOW` compared to `ConceptNet`.

Next we examine the corpus frequency of concepts in `ConceptNet` and `SWOW` using the Google n-gram corpus.[6] Many `ConceptNet` concepts were not present in this large corpus (24%), versus 1.5% for `SWOW`. Of those concepts that could be found, concepts in `SWOW` are on average 7× more common than those in `ConceptNet`. We conclude that `SWOW` concepts are generally common, while `ConceptNet` includes more obscure concepts.

3.2.2 Relational Content

We cannot directly compare relation distributions between `ConceptNet` and `SWOW` because `SWOW` does not have labelled relations. Instead, we inspect the intersection of the two graphs (`CN∩SW`) as all head-tail pairs that are shared between the graphs ($|\mathcal{E}_{\texttt{CN}\cap\texttt{SW}}|$=190K), labelled with their `ConceptNet` relations. We use `CN∩SW` as a proxy of the relations in `SWOW`. For each relation type r, we compare its relative frequency in the full `ConceptNet` ($f^r_{\texttt{CN}}$) against its recall in `CN∩SW` (recall^r), where

$$f^r_{\texttt{CN}} = \frac{|\mathcal{E}_{\texttt{CN}}(r)|}{|\mathcal{E}_{\texttt{CN}}|}, \quad \text{recall}^r = \frac{|\mathcal{E}_{\texttt{CN}\cap\texttt{SW}}(r)|}{|\mathcal{E}_{\texttt{CN}}(r)|}.$$

Figure 3 plots the correlation between the $f^r_{\texttt{CN}}$, and their recall in `CN∩SW`. First, we observe a discenerable correlation between the majority of low- to medium frequency relations in `ConceptNet` and their recall in `CN∩SW` (orange box on bottom left). These relations cover largely semantic associations pertaining to the appearance, use or situational contexts of concepts. Second, none of the six most frequent relation types in `ConceptNet` (right part of the plot), are highly recalled in `CN∩SW`. Out of these, relations indicating (near) synonymy retain a medium recall (red box on middle right), while morphosyntactic relations are less prevalent (blue box on bottom right). These relations are prevalent in `ConceptNet`, because it is derived to a large part from structured, linguistic resources like Wiktionary.[7] Word associations, on the other hand, are known to be dominated by semantic relations (Mollin, 2009). Third, the relations with highest recall in `CN∩SW` tend to be infrequent in `ConceptNet` (green box on top left). Two of these relations focus on *differences* indicating that humans associate contrasting concepts (such as 'hot'→'not cold'; Deese (1964); Clark (1970)). We also found that the proportion of negated edges in `SWOW` (0.7%; N=11K) is more than twice the proportion in `ConceptNet` (0.3%; N=11.5K), and that only 4% of them overlap.[8] Representations of antonyms and negations have traditionally been difficult to infer from text, suggesting that `SWOW` could be a valuable complementary knowledge source to fill this gap.

4 Experiment 2: Coverage of Commonsense Knowledge

This section probes `ConceptNet` and `SWOW` specifically for (situational) *commonsense* knowledge. Daily activities such as *doing the laundry* or *visiting the doctor* involve a wealth of general knowledge touching on causal, temporal, physical, or social knowledge which is rarely explicitly stated (Mostafazadeh et al., 2016; Rashkin et al., 2018; Ostermann et al., 2018, 2019). We leverage the MCScript2.0 data set (Ostermann et al., 2019), a large collection of *explicit* descriptions of everyday scenarios, and investigate whether `ConceptNet` and `SWOW` encode the commonsense knowledge underlying these situations. In

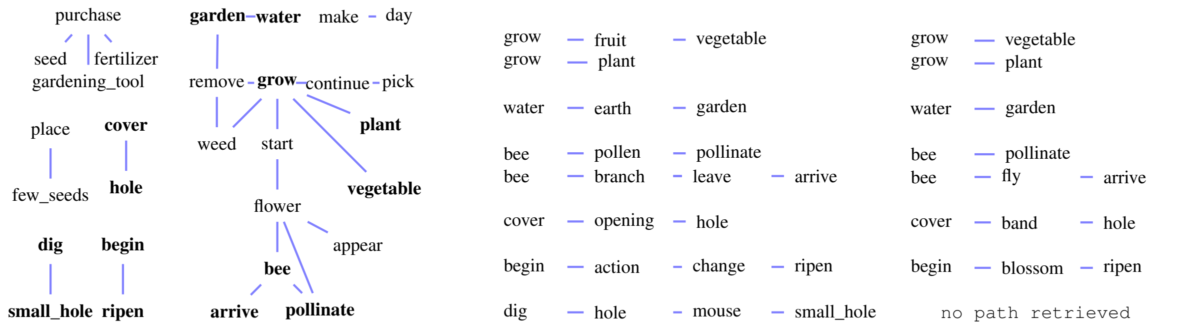
[6] https://books.google.com/ngrams

[7] 74% of `ConceptNet` edges origin from Wiktionary.

[8] Negated nodes were identified based on a list of negation markers, cf., Appendix B.

Growing vegetables In the spring I went to the garden center to **purchase seeds, fertilizer**, and **gardening tools**. Back at home I **dug small holes** into the dirt in my garden, **placed** a **few seeds** in each hole, and **covered** the **holes** with dirt. I then **watered** the **garden**, and **made** sure to water it every **day**. Over the next several weeks I **removed** any **weeds** that were **growing** in the **garden**, and watched as the **plants continued** to **grow**. **Flowers** then **appeared** on the plants, and **bees arrived** to **pollinate** the **flowers**. Then, **started** to **grow** where the **flowers** had been. After a while they **began** to **ripen** and **continued** to **grow** until they were big enough to **pick**.

(a) script text



(b) script graph (c) ConceptNet paths (d) SWOW paths

Figure 4: (a) Narrative describing the scenario of 'growing vegetables' from MCScript2.0; (b) derived SRL graph over predicates ARG0 and ARG1s; (c) shortest parths in ConceptNet for a subset of connected nodes (bolded) in (b); (d) corresponding paths retrieved from SWOW.

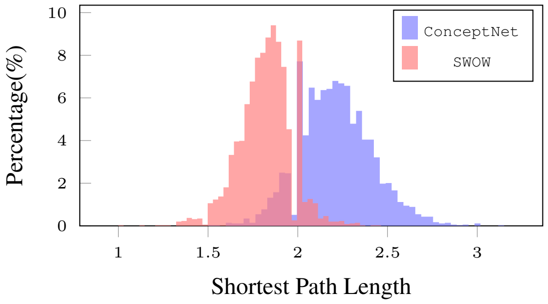


Figure 5: The length distribution of average shortest paths in ConceptNet and SWOW for edges from MCScript graphs.

particular, we test whether the knowledge graph structure underlying MCScript scenarios is retained in ConceptNet and SWOW. Figure 4 shows an example of a scenario (a), derived graph representation (b), and a subset of corresponding associations in ConceptNet (c) and SWOW (d).

4.1 Method

MCScript2.0 is a collection of 3,487 short narrative descriptions covering 200 every-day scenarios of varying complexity (e.g., *cleaning the floor* vs *growing vegetables*) (Ostermann et al., 2019). The descriptions were crowd-sourced, and authors were instructed to describe the underlying scenarios "as if talking to a child" (Ostermann et al., 2018). Thus by design MCScript narratives spell out commonsense knowledge more explicitly than most text corpora. Following prior work on modelling narrative scripts, we posit that narrative chains are fundamentally characterized in terms of their events and participants (Chambers and Jurafsky, 2009; Frermann et al., 2014). We recover this information using semantic role labelling (SRL), and identify predicates, ARG0s and ARG1s in each narrative. We use the resulting set of spans and their relations to transform each narrative (Figure 4a) into a *script graph* (Figure 4b). Spans constitute the nodes of the script graph,[9] and edges correspond to (undirected) SRL dependencies.

Graph mapping We assume that predicates-argument relations in detailed descriptions of everyday events (e.g., 'water' and 'garden') are instances of common sense knowledge, and should hence be directly encoded in a KG. We posit that the shorter the paths, the more directly the necessary scenario-specific commonsense knowledge is encoded in the target KG. E.g., we find a direct 1-hop connection between 'water' and 'garden' in SWOW (Figure 4d), while the path between the concepts in ConceptNet is less direct (2-hops; Figure 4c).

More technically, we project the MCScript graphs onto SWOW and ConceptNet as follows. For all directly connected node pairs in a MCScript graph, we identify their corresponding concepts in

[9] Pronouns and stopwords are excluded. We use the SRL model of Shi and Lin (2019) implemented in AllenNLP.

SWOW and ConceptNet, respectively, by exact string match. In our current study we are primarily interested in *whether* (not how) node pairs from the MCScript graph exist in the target KGs. We therefore treat all graphs as undirected, and retrieve the shortest path between the two nodes in the target KGs. Figure 4c and 4d show a subset of the shortest paths retrieved from ConceptNet and SWOW, respectively, for the MCScript graph in 4b (bolded nodes). See Appendix E for the full list of retrieved paths from ConceptNet and SWOW.

4.2 Results

Figure 5 presents the distribution over shortest path lengths, averaged over the full MCScript data set. We observe that paths, on average, are shorter in SWOW compared to ConceptNet, suggesting that the situational commonsense associations in MCScript are more directly encoded in SWOW. The example shortest paths shown in Figure 4c (ConceptNet) and d (SWOW) further illustrate the associations in the two commonsense resources. The associations imposed in paths of length >1 are meaningful across the board, but differ across the KGs: for example, the connection between 'bee' and 'arrive' is further elaborated in ConceptNet by explaining that in order to arrive, the bee needs to leave (from a plausible location 'branch'); SWOW on the other hand imputes information on the mode of travel ('flying').

As a first exploration into evaluating the extent of commonsense knowledge in commonsense KGs, our method has several weaknesses to be addressed in future work: the string mapping from text to commonsense knowledge could be replaced with more flexible, embedding-based methods; the script graphs themselves could be improved and enriched with more semantic roles, or higher-level narrative structure as captured for instance in Rhetorical Structure Theory (Taboada and Mann, 2006). Similarly, the mapped graphs could incorporate edge directions and/or labels. Ultimately, the mapped paths (most interestingly those mapped to paths of length > 1) will need to be validated and interpreted by human annotators. We believe that leveraging explicit human-created commonsense data sets, like MCScript2.0, opens interesting avenues to understand the commonsense knowledge present in word associations.

5 Experiment 3: Commonsense QA

In this section, we explore the utility of commonsense knowledge in ConceptNet and SWOW in commonsense question answering (CQA) tasks. We incorporate the two KGs into representative and competitive CQA models from the recent literature (Wang et al., 2020; Feng et al., 2020), and apply them to three benchmark data sets. We emphasize that the goal of this study is not competing on leaderboards. Current state-of-the-art models leverage very large language models with billions of parameters (Khashabi et al., 2020), and often draw on additional external resource such as Wiktionary (Xu et al., 2021). Instead, we explore the utility of SWOW and ConceptNet in a selection of representative moderately complex models.

5.1 Experimental setup

Datasets We consider three standard multiple-choice CQA benchmark datasets. **CommonsenseQA** (CSQA; Talmor et al. (2019)) contains commonsense questions generated by crowd workers on the basis of sub-graphs in ConceptNet, giving ConceptNet an inherent advantage over SWOW. The QA-pairs in this dataset require various commonsense skills, and distractor answers were carefully selected to share semantic associations with the key concepts in a question. **OpenBookQA** (OBQA; Mihaylov et al. (2018)) consists of question-answer pairs along with paragraphs from elementary-level science books. Following previous work (Wang et al., 2020; Feng et al., 2020), we disregard the paragraphs, and apply our models to question-answer pairs directly. Building on our analysis in Section 4, we also apply our models to the **MCScript2.0** QA benchmark (Ostermann et al., 2019). Each task consists of a story, and a question paired with two answer options. We use the in-house data split by Lin et al. (2019) for CSQA and the official splits for the other data sets. Table 2 presents data set statistics, as well as an example from each dataset.

Models A typical QA system consists of three modules: (1) a KG encoder which maps a subgraph spanning the concepts in the question/answer to a fixed-dimensional knowledge embedding $\mathbf{k}$; (2) a text encoder, which maps question q and answer a to a fixed-dimensional language embedding $\mathbf{c}$; and (3) a scoring module which scores each answer option given $\mathbf{c}$ and $\mathbf{k}$, and returns the highest scoring answer as a prediction.

Dataset	Example	Train / Dev / Test split
CSQA	What do all humans want to experience in their own home? (a) **feel comfortable**, (b) work hard, (c) fall in love, (d) lay eggs, (e) live forever	8,500/1,221/1,241
OBQA	What is a source of energy? (a) bricks, (b) **grease**, (c) cars, (d) dirt	4,957/500/500
MCScript2.0	When did small plants grow? (a) two days, (b) **after seeds were planted**	14,191/2,020/3,610

Table 2: Details on the benchmarks CSQA, OpenbookQA and MCScript2.0: One example QA-pair per dataset (correct answer in boldface) and sizes of the respective train/dev/test splits. The paragraph of MCScript2.0 example is shown in Figure 4.

Models	CSQA		OBQA		MCScript2.0	
	`ConceptNet`	`SWOW`	`ConceptNet`	`SWOW`	`ConceptNet`	`SWOW`
ALBERT	73.78 (±0.79)		63.47 (±1.42)		93.62 (±0.44)	
+ GconAttn	74.03 (±0.46)	74.05 (±0.50)	65.13 (±2.16)	65.87 (±1.21)	**93.91** (±0.50)	93.84 (±0.35)
+ RN	**75.64** (±0.70)	74.40 (±0.37)	64.73 (±2.10)	**66.40** (±1.00)	93.53 (±0.11)	93.49 (±0.22)

Table 3: Test accuracy on CSQA, OBQA and MCScript2.0. We report performance of ALBERT, and augment it with two KG-aware models using either `ConceptNet` or `SWOW`. Results are averages of the best three out of six runs (based on dev set performance); standard deviations reported in brackets.

We experiment with two KG encoders: GconAttn (Wang et al., 2019) maps question and answer concepts to pre-trained concept embeddings, and then aligns them using concept-level attention and pooling. GconAttn is a relation-free model, leveraging only mentioned concepts from a KG. To disentangle the impact of relation labels, we also experiment with Relation Networks (RN) (Santoro et al., 2017), which embed concepts using context-aware path-level attention over path embeddings. Each path embedding encodes the path between some question concept and answer concept. We use one-hop and two-hop paths in our experiments. We refer the reader to the respective papers for full model details, as well as Appendix D which details our full parameter settings and reproduction results. Both KG encoders require *node embeddings*. We use RoBERTa-Large (Liu et al., 2019) to obtain a node embedding matrix, separately for `ConceptNet` and `SWOW`. Specifically, for each node $\mathcal{E}_i \in \mathcal{E}$, we feed the sequence of [`CLS`] $+ \mathcal{E}_i +$ [`SEP`] to RoBERTa and use the last layer representation of [`CLS`] as its embedding. RN also requires *relation embeddings*. For each of our KGs (`ConceptNet`, `SWOW`), we obtain a separate relation embedding matrix $\mathbf{R}$ with TransE.[10] Following previous work on CQA (Lin et al., 2019; Wang et al., 2020), we select 31 out of `ConceptNet`'s 47 relation types which proved helpful for CQA, and merge the remaining relations into 17 types. We use forward- and mutual associations for `SWOW`. Following previous work (Malaviya et al., 2020; Wang et al., 2020), we densify both `SWOW` and `ConceptNet` by adding for each relation $\mathcal{R}_i$ an additional relation type indicating its reverse $\mathcal{R}_i^{-1}$. For example, a relation $\mathcal{R}_i$=part_of would be complemented with $\mathcal{R}_i^{-1}$=has_part.

We fix the text encoder to ALBERT-xxlarge-v2 (Lan et al., 2020), which performed competitively in recent work (Wang et al., 2020). All models are run six times and we report results using the best three models, as judged by training loss; this method is used to remove outliers resulting from instability of training. All results are our own re-runs using the official implementations from `https://github.com/INK-USC/MHGRN`, and are largely comparable with those reported in the literature. See Appendices C, D for further details.

5.2 Results

Experimental results in Table 3 show that both knowledge-augmented models outperform the language baseline for all data sets except RN on MCScript2.0. The path-aware RN achieves the best performance for both CSQA and OBQA. All models (including the text-only baseline) show comparative and very high performance on MCScript2.0, suggesting that it is a simpler task compared to

[10] Using OpenKE `https://github.com/thunlp/OpenKE`. We do not use TransE node embeddings, as they were outperformed by RoBERTa embeddings in preliminary tests.

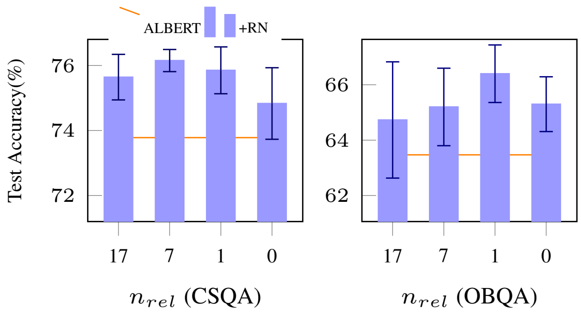


Figure 6: Test set accuracy under different numbers of relation labels for ConceptNet on CSQA (left) and OBQA (right).

the other two. Our models outperform the current state-of-the art on MCScript2.0 by up to 3.24% (absolute).[11] More importantly, models incorporating either of ConceptNet or SWOW achieve similar performance across the board. Recall that CSQA is derived from ConceptNet edges, putting SWOW at a disadvantage. SWOW performs best on OBQA which is independent of both KGs. We measure the significance of differences in performance between the text-only baseline and the RN models on CSQA and OBQA using Student's t-test. We find that the RN models outperform the text-only model significantly (p<0.05) with both ConceptNet and SWOW as underlying KG.[12] There is no significant difference between the ConceptNet-based and the SWOW-based RN models (p>>0.05). These results provide initial evidence that SWOW can be a valuable alternative source of commonsense knowledge to ConceptNet for downstream NLP tasks.

The competitive performance of SWOW may be surprising, particularly with the relation-aware KG encoder RN, because ConceptNet has access to a rich typed relation inventory while SWOW does not. We investigate the impact of labelled relation types on CQA for RN, by ablating the number of relation types accessible to ConceptNet: we grouped the 17 ConceptNet relation types used in the models above into (a) seven coarse types (plus reverse relations) using the relational ontology of Liu and Singh (2004);[13] (b) a single generic relation type (plus reverse relation). This version still contains one-hop and two-hop paths as well as reverse relations; and (c) removing all relation information from the model. Figure 6 shows that merging ConceptNet relations into broader types improves downstream task performance across data sets, with the best results achieved with one or seven relation types. Our results suggest that augmenting SWOW with a rich label inventory may not be necessary for it to be used as a common-sense resource in downstream common-sense reasoning models. Nevertheless, labelling SWOW with cognitively valid relation (or commonsense type) information in order to better understand the types of spontaneous associations humans express is an exciting avenue for future work.

6 Conclusions

We presented an in-depth analysis of the general and commonsense knowledge encoded in human word association norms (SWOW), versus a traditional curated commonsense knowledge graph (ConceptNet). We showed that the two knowledge resources differ systematically in their structure and content. We also showed that SWOW brings comparable gains to ConceptNet when applied to three commonsense question answering benchmarks, which is important as word associations are simpler than structured relations, and accordingly can be created more cheaply via crowdsourcing and without the need for experts. Finally, we showed that SWOW encodes situational commonsense knowledge as encoded in the human-created MCScript2.0 narratives more directly than ConceptNet; and that both KGs impose meaningful additional relations between concepts that were left implicit in the descriptions. There are several directions for future work, most notably extending our framework for characterizing commonsense knowledge in the two KGs, exploring SWOW relation types, developing means of consolidating the two knowledge graphs, and exploring downstream utility of KGs in less well-resourced languages than English, where the portability of the word association annotation methodology confers a substantial advantage.

Acknowledgements

This work was supported in part by China Scholarship Council (CSC). We thank the reviewers for their valuable comments, and Simon De Deyne for insightful discussions.

[11] https://coinnlp.github.io/task1.html

[12] The only exception is ALBERT vs RN with ConceptNet on OBQA where p=0.07.

[13] See Appendix A for details of 17 and 7 relation types.

References

Maria A. Rodriguez and Paola Merlo. 2020. Word associations and the distance properties of context-aware word embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*.

Javier Borge-Holthoefer and Alex Arenas. 2010. Categorizing words through semantic memory navigation. *The European physical journal B*, 74(2):265–270.

Erik Cambria, Yang Li, Frank Z. Xing, Soujanya Poria, and Kenneth Kwok. 2020. Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

Herbert H. Clark. 1970. Word associations and linguistic theory. *J. Lyons (Ed.), New horizons in linguistics*, 3:271–286.

Jeff Da and Jungo Kasai. 2019. Cracking the contextual commonsense code: Understanding commonsense reasoning aptitude of deep contextual representations. 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)*.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58:92–103.

Simon De Deyne, D Navarro, and Gert Storms. 2013. Associative strength and semantic activation in the mental lexicon: evidence from continued word associations. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*.

Simon De Deyne, Amy Perfors, and Daniel J Navarro. 2016. Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*.

J. Deese. 1964. The associative structure of some common english adjectives. *Journal of Verbal Learning and Verbal Behavior*, 3:347–357.

J. Deese. 1966. The structure of associations in language and thought.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*.

Simon De Deyne, Danielle J. Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The “Small World of Words” English word association norms for over 12,000 cue words. *Behavior Research Methods*, 51:987–1006.

Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021. Back to square one: Bias detection, training and commonsense disentanglement in the winograd schema. *arXiv preprint arXiv:2104.08161*.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

T. Fitzpatrick. 2006. Habits and rabbits: word associations and the l2 lexicon. *Eurosla Yearbook*, 6:121–145.

Lea Frermann, Ivan Titov, and Manfred Pinkal. 2014. A hierarchical bayesian model for unsupervised induction of script knowledge. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*.

Andrew S Gordon and Jerry R Hobbs. 2017. *A formal theory of commonsense psychology: How people think people think*. Cambridge University Press.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *In Proceedings of the 2013 workshop on Automated knowledge base construction*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6384–6392.

Filip Ilievski, A. Oltramari, Kaixin Ma, Bin Zhang, D. McGuinness, and Pedro A. Szekely. 2021a. Dimensions of commonsense knowledge. *ArXiv*, abs/2101.04640.

Filip Ilievski, Pedro Szekely, Jingwei Cheng, Fu Zhang, and Ehsan Qasemi. 2020. Consolidating commonsense knowledge. *arXiv preprint arXiv:2006.06114*.

Filip Ilievski, Pedro A. Szekely, and B. Zhang. 2021b. Cskg: The commonsense knowledge graph. In *ESCW*.

Jaeyoung Jung, Na Li, and Hiroyuki Akama. 2010. Network analysis of Korean word associations. In *Proceedings of the First Workshop on Computational Neurolinguistics*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, P. Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *ArXiv*, abs/2005.00700.

G. R. Kiss, C. Armstrong, R. Milroy, and J. Piper. 1973. An associative thesaurus of english and its computer analysis. *The Computer and Literary Studies*.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR*.

DB Lenat and RV Guha. 1993. Building large knowledge-based systems: Representation and inference in the cyc project. *Artificial Intelligence*, 61(1):4152.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. 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)*.

Hugo Liu and Push Singh. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22:211–226.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2020. Exploiting structural and semantic context for commonsense knowledge base completion. *The Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. 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*.

Sandra Mollin. 2009. Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus Linguistics & Linguistic Theory*, 5(2).

N. Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, P. Kohli, and James F. Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

D. Nelson, C. McEvoy, and S. Dennis. 2000. What is free association and what does it measure? *Memory & Cognition*, 28:887–899.

Douglas L. Nelson, Cathy McEvoy, and Thomas A. Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36:402–407.

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018. MCScript: A novel dataset for assessing machine comprehension using script knowledge. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Simon Ostermann, Michael Roth, and Manfred Pinkal. 2019. MCScript2.0: A machine comprehension corpus focused on script events and participants. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 103–117, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay Pujara, Eriq Augustine, and Lise Getoor. 2017. Sparsity and Noise: Where Knowledge Graph Embeddings Fall Short. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI preprint*.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling Naive Psychology of Characters in Simple Commonsense Stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Tara Safavi and Danai Koutra. 2021. Relational world knowledge representation in contextual language models: A review. *ArXiv*, abs/2104.05837.

Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Tim Lillicrap. 2017. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence*.

Peng Shi and Jimmy Lin. 2019. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *ArXiv*, abs/1904.05255.

Daniel J. Navarro Simon De Deyne and Gert Storms. 2013. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods*, 45:480–498.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

Maite Taboada and William C Mann. 2006. Rhetorical structure theory: Looking back and moving ahead. *Discourse studies*, 8(3):423–459.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Hendrik Vankrunkelsven, Steven Verheyen, Gert Storms, and Simon De Deyne. 2018. Predicting lexical norms: A comparison between a word association model and text-based word co-occurrence models. *Journal of cognition*, 1(1).

Peifeng Wang, Nanyun Peng, Filip Ilievski, Pedro Szekely, and Xiang Ren. 2020. Connecting the dots: A knowledgeable path generator for commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, and Michael J. Witbrock. 2019. Improving natural language inference using external knowledge in the science questions domain. In *The Thirty-Third AAAI Conference on Artificial Intelligence*.

Yichong Xu, Chenguang Zhu, Ruochen Xu, Yang Liu, Michael Zeng, and Xuedong Huang. 2021. Fusing context into knowledge graph for commonsense question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1201–1207.

A Relation types list

For ConceptNet, we start with grouping the 31 original ConceptNet relation types into 17 clusters following (Wang et al., 2020) (Table 4 left). We further group the 31 relation types into 7 by following (Liu and Singh, 2004), which grouping relation types into 7 categories, including {things, spatial, events, causal, affective, functional, agents} (Table 4 right).

B Negation List

We use a list of negation markers to identify negated words and phrases. The list of negation markers is: ('no', 'not', 'none', 'nor', 'no one', 'nobody', 'nothing', 'neither', 'nowhere', 'never', 'hardly', 'barely', 'scarcely', 'non', 'without', 'fail', 'cannot', 'cant', 'no longer', 'dont', 'wont'). Some negated triples sampled from ConceptNet ans SWOW are presented in Table 5.

C Hyper-parameters

We use the cross-entropy as the loss function with RAdam (Liu et al., 2020) as optimizer to train all the models. We use as GELU (Hendrycks and Gimpel, 2016) as the activation function. We report most important hyper-parameters in Table 6.

D Re-Production Results

For all KG-augmented models, we use the implementation from previous work (Wang et al., 2020). [14] Table 7 compares our re-run results to the original numbers reported in the respective papers. All our reproduced scores are comparable to or better than reported numbers. All of our experiments are run on single GPU of NVIDIA V100 16G.

E Full list of shortest paths

Table 8 presents all directed connected paths shown in Table 4 and their corresponding shortest paths retrieved ConceptNet and SWOW.

[14] https://github.com/INK-USC/MHGRN

17 relation types			
1	atlocation, locatednear	10	usedfor
2	capableof	11	receivesaction
3	createdby	12	madeof
4	desires	13	partof, hasa
5	hascontext	14	notdesires
6	hasproperty	15	notcapableof
7	antonym, distinctfrom	16	isa, instanceof, definedas
8	relatedto, similarto, synonym	17	causes, causesdesire, motivatedbygoal
9	hassubevent, hasfirstsubevent, haslastsubevent, hasprerequisite, entails, mannerof		

7 relation types	
1	capableof
2	usedfor, receivesaction
3	atlocation, locatednear, hascontext, similarto
4	causes, causesdesire, motivatedbygoal, desires
5	antonym, distinctfrom, notcapableof, notdesires
6	isa, hasproperty, madeof, partof, definedas, instanceof, hasa, createdby, relatedto, synonym
7	hassubevent, hasfirstsubevent, haslastsubevent, hasprerequisite, entails, mannerof

Table 4: Conflation of relation types in ConceptNet to either 17 (left) or 7 (right) coarser grained groups.

ConceptNet negated triples	SWOW negated triples
(antonym, still with us, no longer with us)	(forwardassociated, love, non tangible)
(antonym, awesome, fail)	(forwardassociated, slight, barely)
(antonym, able, cannot)	(forwardassociated, real, not fake)
(relatedto, nobody, no one)	(forwardassociated, tedious, not fun)
(synonym, zero, nothing)	(forwardassociated, everything, nothing)
(antonym, both, neither)	(forwardassociated, with, without)
(capableof, clues, lead nowhere)	(forwardassociated, broke, no money)
(causes, going to sleep, never waking up)	(forwardassociated, lemon, not lime)
(distinctfrom, hardly, easily)	(forwardassociated, later, not now)
(hassubevent, eat quickly, barely chew)	(forwardassociated, punishment, not good)
(derivedfrom, scarcely, scarce)	(forwardassociated, none, no more)
(causes, dying, non existence)	(forwardassociated, succeed, fail)
(hassubevent, eat healthily, dont eat junk food)	(forwardassociated, unable, cannot)
(relatedto, dare, you wont)	(forwardassociated, obsolete, no longer exists)

Table 5: Examples of negated triples from ConceptNet and SWOW.

Type	Hyperparameter	Value
General		
	batch size	32/16/16
	dropout	0.1/0.2/0.1
	early stopping patience	2 epochs
	max sentence length	80/84/300
	weight decay	0.01
ALBERT-xxlarge-v2		
	learning rate	1e-05
GconAttn		
	learning rate	3e-04/3e-04/1e-03
	MLP layers	2
	hidden units	{256, 128}
	concept embedding dimension	1024
RN		
	learning rate	1e-03/3e-04/1e-03
	MLP layers	3
	hidden units	{256, 256, 128}
	concept embedding dimension	1024
	relation embedding dimension	100

Table 6: Hyperparameters for various models and data sets. Values split by “/" follow the order of CSQA/OBQA/MCScript2.0.

Model	CSQA		OBQA	
	(Wang et al., 2020)	Our re-run	(Wang et al., 2020)	Our re-run
w/o KG	68.69 (±0.56)	70.46 (±0.18)	64.80 (±2.37)	64.47 (±3.01)
+ GconAttn	69.88 (±0.47)	70.59 (±0.66)	64.75 (±1.48)	69.00 (±1.41)
+ RN	69.59 (±3.80)	72.79 (±0.63)	65.20 (±1.18)	65.30 (±0.99)

Table 7: Comparisons of our re-production of various KG-augmented models with previous work. The RoberTa-large encoder is used. We use the same provided code by (Wang et al., 2020) for CSQA and OBQA. Note that text representation on OBQA is the the average pooling over the hidden states of the last layer of RoberTa rather than 'CLS' token representation.

	MCScript	ConceptNet	SWOW
1	(purchase, seed)	(purchase, sale, full, seed)	(purchase, need, seed)
2	(purchase, fertilizer)	(purchase, chain, garage, fertilizer)	(purchase, product, fertilizer)
3	(hole, cover)	(hole, opening, cover)	(hole, band, cover)
4	(flower, appear)	(flower, visit, appear)	(flower, become, appear)
5	(flower, pollinate)	(flower, pollen, pollinate)	(flower, bee, pollinate)
6	(flower, start)	(flower, open, start)	(flower, green, start)
7	(flower, bee)	(flower, bee)	(flower, bee)
8	(grow, continue)	(grow, carry, continue)	(grow, extend, continue)
9	(grow, vegetable)	(grow, fruit, vegetable)	(grow, vegetable)
10	(grow, plant)	(grow, plant)	(grow, plant)
11	(grow, weed)	(grow, field, weed)	(grow, weed)
12	(continue, pick)	(continue, carry, pick)	(continue, stick, pick)
13	(pollinate, bee)	(pollinate, pollen, bee)	(pollinate, bee)
14	(ripen, begin)	(ripen, change, action, begin)	(ripen, blossom, begin)
15	(garden, water)	(garden, earth, water)	(garden, water)
16	(garden, remove)	(garden, cricket, remove)	(garden, wart, remove)
17	(weed, remove)	(weed, remove)	(weed, remove)
18	(small hole, dig)	(small hole, mouse, hole, dig)	-
19	(bee, arrive)	(bee, branch, leave, arrive)	(bee, fly, arrive)
20	(day, make)	(day, clear, make)	(day, present, make)
21	(few seed, place)	-	-
22	(gardening tool, purchase)	(gardening tool, tool, lever, purchase)	-

Table 8: Full list of paths from ConceptNet and SWOW for example shown in Figure 4. - indicates path are not retrieved in the target KG.

.