Dimensions of Commonsense Knowledge

Filip Ilievski[a,*], Alessandro Oltramari[b], Kaixin Ma[c], Bin Zhang[a], Deborah L. McGuinness[d] and Pedro Szekely[a]

[a]*Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA*

[b]*Intelligent Internet of Things, Bosch Research and Technology Center, Pittsburgh, PA, USA*

[c]*Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA*

[d]*Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, USA*

ARTICLE INFO

Keywords:
commonsense knowledge
semantics
knowledge graphs
reasoning

ABSTRACT

Commonsense knowledge is essential for many AI applications, including those in natural language processing, visual processing, and planning. Consequently, many sources that include commonsense knowledge have been designed and constructed over the past decades. Recently, the focus has been on large text-based sources, which facilitate easier integration with neural (language) models and application to textual tasks, typically at the expense of the semantics of the sources and their harmonization. Efforts to consolidate commonsense knowledge have yielded partial success, with no clear path towards a comprehensive solution. We aim to organize these sources around a common set of dimensions of commonsense knowledge. We survey a wide range of popular commonsense sources with a special focus on their relations. We consolidate these relations into 13 knowledge dimensions. This consolidation allows us to unify the separate sources and to compute indications of their coverage, overlap, and gaps with respect to the knowledge dimensions. Moreover, we analyze the impact of each dimension on downstream reasoning tasks that require commonsense knowledge, observing that the temporal and desire/goal dimensions are very beneficial for reasoning on current downstream tasks, while distinctness and lexical knowledge have little impact. These results reveal preferences for some dimensions in current evaluation, and potential neglect of others.

1. Introduction

Commonsense knowledge is information that humans typically have that helps them make sense of everyday situations. As such, this knowledge can generally be assumed to be possessed by most people, and, according to the Gricean maxims [27], it is typically omitted in (written or oral) communication. The fact that commonsense knowledge is often implicit presents a challenge for automated natural language processing (NLP) and question answering (QA) approaches as the extraction and learning algorithms cannot count on the commonsense knowledge being available directly in text.

Due to its prominence and implicit nature, capturing commonsense knowledge holds a promise to benefit various AI applications, including those in NLP, computer vision, and planning. For instance, commonsense knowledge can be used to fill gaps and explain the predictions of a (neural) model [47], understand agent goals and causality in stories [88], or enhance robot navigation and manipulation [91]. Its benefits have already been shown in downstream applications, such as software requirements localization [63] and aligning local laws with public reactions [66].

Consequently, acquiring and representing commonsense knowledge in a machine-readable form, as well as reasoning with it, has been a major pursuit of AI since its early days [55]. This has manifested in the design, construction, and curation of a rich palette of resources that include com-

monsense information (potentially along with other content) like Cyc [43], ATOMIC [72], WebChild [84], ConceptNet [81], WordNet [56], FrameNet [3], and Visual Genome [40]. Some of these, such as ConceptNet and Cyc, have been deliberately created to capture information that would be useful for common sense-related reasoning tasks, while others, like WordNet or Visual Genome, were intended to support other tasks such as word sense disambiguation or image object recognition. As reported in [35], the commonsense sources exhibit great diversity in terms of their representational formats, creation methods, and coverage. While there is an opportunity for this knowledge to be exploited jointly, the inherent diversity makes the consolidation of these sources challenging.

Meanwhile, the last few years have featured an increased focus on benchmarks that can be used to evaluate different aspects of common sense, including social [73], physical [8], visual [92], and numeric [48] common sense. Further distinction has been made between discriminative tasks [73, 8, 83], where the goal is to pick the single correct answer from a list, and generative tasks, where one has to generate one or multiple correct answers [48, 9]. These tasks can be tackled by using (the entire or a subset of the) training data [52, 47], or in a zero-/few-shot evaluation regime [53, 78].

The wealth and diversity of commonsense sources, on one hand, and benchmarks, on the other hand, raises a natural question: what is the role of these knowledge repositories in real-world reasoning techniques that need to incorporate commonsense knowledge? While intuitively such sources of commonsense knowledge can have tremendous value for reasoning in downstream reasoning tasks, the practice shows that their impact on these tasks has been relatively limited, especially in comparison to the contribution of the

*Corresponding author
✉ ilievski@isi.edu (F. Ilievski); Alessandro.Oltramari@us.bosch.com (A. Oltramari); kaixinm@andrew.cmu.edu (K. Ma); binzhang@isi.edu (B. Zhang); dlm@cs.rpi.edu (D.L. McGuinness); pszekely@isi.edu (P. Szekely)
ORCID(s):

language models (LMs).[1] Knowledge sources tend to have larger contributions when little or no training data is available: for example, one can generate artificial training sets based on several sources, which can be used to pre-train language models and apply them on downstream tasks without using the official training data [4, 53]. The impact of knowledge resources so far has been generally restricted to the special cases where the knowledge and the task are known (in advance) to be well-aligned [53, 52]. While a variety of sources [72, 56, 81] or their combination [35] have been used to enhance language models for downstream reasoning, little is known about how this alignment between knowledge types and tasks can be dynamically achieved.

Most recent sources have focused on the breadth of knowledge, sometimes at the expense of its semantics [6, 60]. Text-based representations are particularly attractive, as they facilitate a more direct integration with language models, as well as reasoning on NLP and QA tasks. These sources are often treated as 'corpora', where each fact is typically lexicalized (manually or automatically) into a single sentence [52], which is used to inform or fine-tune a language model. Due to the lack of focus on formal representational principles, the sources capture knowledge types which are not trivial to align with other sources, as shown by the sparse mappings available between these sources [35]. Considering the lack of a common vocabulary and the lack of alignment of these sources, their limited coverage, and lack of focus on explicit semantics, knowledge is typically kept in an impoverished textual form that is easy to capture and combine with language models. The downsides of this practice are: 1) the commonsense knowledge remains difficult to harmonize across sources; 2) without a thorough harmonization or consolidation, it is not clear how to effectively measure coverage, overlap, or gaps; and 3) text-based representations alone may be unable to match the human-like level of sophistication typical for contextual reasoning.

Efforts to consolidate commonsense knowledge across sources [35, 22, 61] have managed to bring these sources closer, which has shown impact on commonsense QA tasks [53]. In [34], we provide heuristics for defining the boundaries of commonsense knowledge, in order to extract such a subset from one of the largest available graphs today, Wikidata [87]. Yet, these efforts have limited success, and many consolidation questions are left open. How should one think about commonsense knowledge in a theoretical way? What does it mean to build a consolidated knowledge graph (KG) of resources created largely in a bottom-up fashion? How should the relations be chosen? What is the right level of abstraction for relations and nodes?

2. Approach

We aim to provide insights into such questions, aiming primarily to organize the types of knowledge found in existing sources of commonsense knowledge.

For this purpose, we survey a wide variety of sources of commonsense knowledge, ranging from commonsense KGs through lexical and visual sources, to the recent idea of using language models or corpora as commonsense knowledge bases. We survey their relations and group them into a set of **dimensions**, each being a cluster of its specific relations, as found in the sources. We then apply these dimensions to transform and unify existing sources, providing an enriched version of the Commonsense Knowledge Graph (CSKG) [35]. The dimensions allow us to perform four novel experiments:

1. We assess the coverage of the sources with respect to each dimension, noting that some sources have wide (but potentially shallow) coverage of dimensions, whereas others have deep but narrow coverage. This supports the need to integrate these complementary sources into a single one.
2. The consolidation of the dimensions enables us to compare knowledge between the sources and compute metrics of overlap. The results show that there is little knowledge overlap across sources, even after consolidating the relations according to our dimensions.
3. The statements that are expressed with a dimension form a cluster of knowledge. We contrast these dimension-based clusters to language model-based clusters, which allows us to understand some similarities and differences in terms of their focus.
4. We measure the impact of each dimension on two representative commonsense QA benchmarks. Following [53], we pre-train a language model and apply it on these benchmarks in a zero-shot fashion (without making use of the task training data). The dimensions provide a more direct alignment between commonsense knowledge and the tasks, revealing that some dimensions of knowledge are far more informative for reasoning on these tasks than others.

The contributions of the paper are as follows. 1) We survey a wide variety of existing sources of commonsense knowledge, with an emphasis on their relations. We provide a categorization of those sources and include a short overview of their focus and creation methods (**Section 3**). 2) We analyze the entire set of relations and abstract them to a set of 13 commonsense dimensions. Each dimension abstracts over more specific relations, as found in the sources (**Section 4**). 3) The identified dimensions are applied to consolidate the knowledge in CSKG, which integrates seven of the sources we analyze in this paper. The resulting resource is publicly available (**Section 5**). 4) We make use of this dimension-based enrichment of CSKG to analyze the overlap, coverage, and knowledge gaps of individual knowledge sources in CSKG, motivating their further consolidation into a single resource (**Sections 5.1 - 5.3**). 5) We evaluate the impact of different dimensions on two popular downstream commonsense reasoning tasks. The results show that certain dimensions, like temporal and motivational knowledge, are very beneficial and well-covered by benchmarks, whereas

---

[1]This phenomenon can be seen on the benchmark leadearboards, which are dominated by 'pure' language models, for instance: https://leaderboard.allenai.org/socialiqa/submissions/public (accessed on January 5th, 2021).

Table 1

Overview of commonsense knowledge sources. The asterisk ('*') indicates that the source is extended with WordNet knowledge. For FrameNet and MetaNet, we specify their numbers of frame-to-frame relations. WebChild contains a large number of relations, expressed as WordNet synsets, which are aggregated into 4 groups.

Category	Source	Relations	Example 1	Example 2
Commonsense KGs	ConceptNet*	34	*food - capable of - go rotten*	*eating - is used for - nourishment*
	ATOMIC	9	*Person X bakes bread - xEffect - eat food*	*PersonX is eating dinner - xEffect - satisfies hunger*
	GLUCOSE	10	*Someone$_A$ makes Something$_A$ (that is food)*	*Causes/Enables Someone$_A$ eats Something$_A$*
	WebChild	4 (groups)	*restaurant food - quality#n#1 - expensive*	*eating - type of - consumption*
	Quasimodo	78,636	*pressure cooker - cook faster - food*	*herbivore - eat - plants*
	SenticNet	1	*cold_food - polarity - negative*	*eating breakfast - polarity - positive*
	HasPartKB	1	*dairy food - has part - vitamin*	*n/a*
	Probase	1	*apple - is a - food*	*n/a*
	Isacore	1	*snack food - is a - food*	*n/a*
Common KGs	Wikidata	6.7k	*food - has quality - mouthfeel*	*eating - subclass of - ingestion*
	YAGO4	116	*banana chip - rdf:type - food*	*eating - rdfs:label - feeding*
	DOLCE*	1	*n/a*	*n/a*
	SUMO*	1,614	*food - hyponym - food_product*	*process - subsumes - eating*
Lexical resources	WordNet	10	*food - hyponym - comfort food*	*eating - part-meronym - chewing*
	Roget	2	*dish - synonym - food*	*eating - synonym - feeding*
	FrameNet	8 (f2f)	*Cooking_creation - has frame element - Produced_food*	*eating - evoke - Ingestion*
	MetaNet	14 (f2f)	*Food - has role - food_consumer*	*consuming_resources - is - eating*
	VerbNet	36 (roles)	*feed.v.01 - Arg1-PPT - food*	*eating - hasPatient - comestible*
Visual sources	Visual Genome	42,374	*food - on - plate*	*boy - is eating - treat*
	Flickr30k	1	*a food buffet - corefers with - a food counter*	*a eating place - corefers with - their kitchen*
Corpora & LMs	GenericsKB	n/a	*Aardvarks search for food.*	*Animals receive nitrogen by eating plants.*
	GPT-2	n/a	*Food causes a person to be hungry and a person to eat.*	*Eating at home will not lead to weight gain.*

other dimensions like distinctness and lexical knowledge currently have little impact. These results show that our dimensions can be used to facilitate more precise alignment between the available knowledge and question answering tasks. This evaluation also points to gaps in the current knowledge sources and tasks (**Section 5.4**). 6) We reflect on the results of our analysis, and use the resulting observations as a basis to provide a roadmap towards building a more comprehensive resource that may further advance the representation of, and reasoning with, commonsense knowledge. Such a resource would be instrumental in building a general commonsense service in the future (**Section 7**).

3. Sources of Commonsense Knowledge

We define a **digital commonsense knowledge source** as a potentially multi-modal repository from which commonsense knowledge can be extracted.[2] Commonsense knowledge sources come in various forms and cover different types of knowledge. While only a handful of sources have been formally proposed as commonsense sources, many others cover aspects of common sense. Here, we collect a representative set of sources, which have been either proposed as, or considered as, repositories of commonsense knowledge in the past. We categorize them into five groups, and describe the content and creation method of representative sources within each group. Table 3 contains statistics and examples for each source.

---

[2]For brevity, we omit the word 'digital' in the remainder of this paper.

3.1. Commonsense Knowledge Graphs

ConceptNet [81] is a multilingual commonsense knowledge graph. Its nodes are primarily lexical and connect to each other with 34 relations. Its data is largely derived from the crowdsourced Open Mind Common Sense corpus [80], and is complemented with knowledge from other resources, like WordNet.

ATOMIC [72] is a commonsense knowledge graph that expresses pre- and post-states for events and their participants in a lexical form with nine relations. Its base events are collected from a variety of corpora, while the data for the events is collected by crowdsourcing.

GLUCOSE [60] contains causal knowledge through 10 relations about events, states, motivations, and emotions. The knowledge in GLUCOSE is crowdsourced based on semi-automatic templates, and generalized from individual stories to more abstract rules.

WebChild [84] is a commonsense knowledge graph whose nodes and relations are disambiguated as WordNet senses. It captures 20 main relations, grouped in four categories. WebChild has been extracted automatically from Web information, and then is canonicalized in a post-processing step.

Quasimodo [71] contains commonsense knowledge about object properties, human behavior, and general concepts, expressed with nearly 80K relations. Its nodes and relations are initially lexical and extracted automatically from search logs and forums, after which a notable subset of them has been clustered into WordNet domains.

SenticNet [12] is a knowledge base with conceptual and affective knowledge, which is extracted from text and aggregated automatically into higher-level primitives.

HasPartKB [7] is a knowledge graph of hasPart state-

ments, extracted from a corpus with sentences and refined by automatic means.

Probase [90] is a probabilistic taxonomy of IsA statements, which has been automatically extracted from a large corpus.

IsaCore [13] is a taxonomy of IsA knowledge selected from ConceptNet and Probase.

3.2. Common Knowledge Graphs and Ontologies

Wikidata [87] is a general-domain knowledge graph, tightly coupled with Wikipedia, that describes notable entities. Its nodes and relations are disambiguated as Qnodes. The content of Wikidata is collaboratively created by humans, as well as other existing sources. Given the vast number of statements in Wikidata and its sizable set of over 7 thousand relations, we consider its *Wikidata-CS* commonsense subset, as extracted by Ilievski et al. [34].

YAGO [86] is a general-purpose knowledge graph, whose nodes and relations are disambiguated entities. YAGO models its knowledge with 116 relations. The knowledge in YAGO is extracted automatically from Wikipedia, and consolidated with knowledge from other sources, like Schema.org [28].

DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) [23] is an upper level ontology that captures the ontological categories underlying natural language and human common sense with disambiguated concepts and relations. It has been created manually by experts.

SUMO (Suggested Upper Merged Ontology) [62] is an ontology of upper-level disambiguated concepts and their relations. It has been created manually by experts.

3.3. Lexical Resources

WordNet [56] is a lexical database of words, their meanings, and taxonomical organization, in over 200 languages. WordNet has nearly a dozen relations. It has been created manually by experts.

Roget [69] is a manually-created thesaurus that contains synonyms and antonyms for English words.

FrameNet [3] is a lexical resource that formalizes the frame semantics theory: meanings are mostly understood within a frame of an event and its participants that fulfill roles in that frame. FrameNet has 8 frame-to-frame relations. FrameNet was created manually by experts.

MetaNet [18] is a repository of conceptual frames, as well as their relations which often express metaphors. There are 14 frame-to-frame relations in MetaNet. It has been created manually.

VerbNet [75] is a resource that describes syntactic and semantic patterns of verbs, and organizes them into verb classes. There are 36 relations about roles in VerbNet. VerbNet has been created manually by experts.

3.4. Visual Commonsense Sources

Visual Genome [40] contains annotations of concepts and their relations in a collection of images. VisualGenome has over 40K relations. The image descriptions are manually written by crowd workers, while their concepts are mapped

automatically to WordNet senses and revised by crowd workers.

Flickr30k [65] annotates objects in 30k images by multiple workers. The expressions used by different annotators are clustered automatically into groups of coreferential expressions by van Miltenburg [58].

3.5. Corpora and Language Models

GenericsKB [6] contains self-contained generic facts represented as naturally occurring sentences. The sentences have been extracted from three existing corpora, filtered by handwritten rules, and scored with a BERT-based classifier.

Language models, like RoBERTa [51] and GPT-2 [67], can be used as KGs [64] to complement explicitly stated information, e.g., as a link prediction system like COMET [10] or through self-talk [78].

3.6. Observations

As apparent from the descriptions in this section, the existing commonsense sources are based on a wide range of representation principles and have been created using different construction methods. Through the example scenarios of food and eating (Table 3), we show that they have notable overlap in terms of their covered (typically well-known) concepts. At the same time, the types of knowledge covered differ across sources: some sources provide truisms, such as feeding is done with food, while others speculate on usual properties of food, such as its capability to spoil (or "go rotten") or often be placed on a plate. Furthermore, we observe that same or similar relations tend to have different names across sources (compare *type of* to *subclass of* or *is*; or *has quality* in Wikidata to *cook faster* in Quasimodo).

These distinctions make the integration of these sources, and the understanding of their coverage and gaps, very challenging. In order to integrate the knowledge in these sources, we next propose a consolidation of their relations into a common set of dimensions.

4. Dimensions of commonsense knowledge

In the previous section, we surveyed 20 representative commonsense sources from five categories: commonsense KGs, common KGs, lexical sources, visual sources, and corpora and language models. A key contribution of this paper is a manual categorization (by the authors) of the kind of knowledge expressed by the relations in these sources into 13 dimensions. Table 2 shows the correspondence of each relation in these analyzed sources to our dimensions. An example for each of the dimensions from different sources is shown in Table 4. We next describe each dimension in turn.

Lexical. Many data sources leverage the vocabulary of a language or the lexicon in their relations. This includes relationships such as plural forms of nouns, or past tenses of verbs, for example. Lexical knowledge also covers substring information. ConceptNet, for example, includes a relationship called `DerivedFrom` that they describe as capturing when a word or phrase appears within another term and contributes to that term's meaning. Lexical knowledge is also

Table 2
Knowledge dimensions with their corresponding relations. Relations marked with '*' describe knowledge in multiple dimensions. Relations with ¬ express negated statements. Sources in the column *Other*: FN = FrameNet, WN = WordNet, RG = Roget, HP = HasPartKB, PB = Probase, IC = IsaCore, SN = SenticNet.

Dimension	ATOMIC	ConceptNet	WebChild	Other	Wikidata
lexical		FormOf DerivedFrom EtymologicallyDerivedFrom		lexical_unit (FN) lemma (WN)	label
similarity		Synonym SimilarTo DefinedAs	hassimilar	reframing_mapping (FN) metaphor (FN) Synonym (RG) synonym (WN)	said to be the same as
distinctness		Antonym DistinctFrom		Antonym (RG) antonym (WN) excludes (FN)	different from opposite of
taxonomic		IsA InstanceOf MannerOf	hasHypernymy	perspective_on (FN) inheritance (FN) hypernym (WN) isA (PB, IC)	subClassOf instanceOf description
part-whole		PartOf HasA MadeOf AtLocation*	physicalPartOf memberOf substanceOf	HasPart (HP) meronym (WN) holonym (WN)	has part member of material used
spatial		AtLocation* LocatedNear	location spatial		location anatomical location
creation		CreatedBy			creator
utility		ReceivesAction UsedFor CapableOf ¬NotCapableOf	hassynsetmember activity participant	using (FN)	used by use uses
desire/goal	xIntent xWant oWant	CausesDesire MotivatedByGoal Desires ¬NotDesires ObstructedBy			
quality	 xAttr	 HasProperty ¬NotHasProperty SymbolOf	shape size color taste_property temperature	 frame_element (FN) polarity (SN)	 color has quality
comparative			*6.3k relations*		
temporal	xNeed xEffect oEffect xReact oReact	HasFirstSubevent HasLastSubevent HasSubevent HasPrerequisite Causes Entails	time emotion prev next	subframe (FN) precedes (FN) inchoative_of (FN) causative_of (FN)	 has cause has effect
relational -other		RelatedTo HasContext EtymologicallyRelatedTo	thing agent	see_also (FN) requires (FN)	field of this occupation depicts health specialty

the formalization of the relation between a concept and its expression in a language, e.g., denoted by through the `label` relation in Wikidata.

Similarity. Most data sources include the notion of synonymy between expressions. They typically allow definitions of terms, and some cover a broader notion of general similarity. ConceptNet has all three subcategories - for instance, regarding similarity, it establishes that wholesome and organic food are similar notions, while eating is defined as process of taking in food. WebChild also captures sim-

ilarity between WordNet concepts, while WordNet, Wikidata, and Roget focus on synonymy. For instance, Roget declares that food and edibles are synonyms, while Wikidata expresses that food is *said to be the same as* nutriment.

Distinctness. Complementary to similarity, most data sources have notions of some kind of distinguishability. Most commonly, this is formalized as antonymy, where words have an opposition relationship between them, i.e., they have an inherently incompatible relationship. For example, both Roget and ConceptNet consider hot and cold to be antonyms, as

these are two exclusive temperature states of objects. FrameNet defines an `Excludes` relation to indicate that two roles of a frame cannot be simultaneously filled in a given situation. For instance, in the *Placing* frame, an event can either be brought by a cause event or by an intentional agent, but not both. Weaker forms of distinctness are defined by Wikidata and ConceptNet, for concepts that might be mistaken as synonyms. For example, Wikidata states that food safety is different from food security, while ConceptNet distinguishes food from drinks.

Taxonomic. Most data sources include a kind of arrangement classification where some objects are placed into more general and more specific groupings with inheritance relations. When those groupings are ordered categories based on generality, this captures the notion of hyponymy, indicating a subcategory relationship. Hyponymy blends the distinction between the relationships `subclass` (intended for two classes) and `InstanceOf` (intended as a relation between an instance and a class). For instance, Wikidata states that a sandwich wrap is street food, and that food is a disposable product. WordNet has information that beverage and comfort food are hyponyms of food. While this dimension generally focuses on concepts (nouns), it also includes a specialization relation for verbs. Here, the `MannerOf` relation in ConceptNet states that wheezing is a manner of breathing. Two of the sources we review in this paper, Probase [90] and IsaCore [13], focus only on taxonomic relations.

Part-whole. Many data sources include a notion of being a part of or a member of something. Part-whole knowledge can be transitive, such as that of geographic containment, exemplified by New York City being a part of New York State, which is also part of the United states. Other part-of notions, such as *member-of* are not necessarily transitive. A third category of part-whole knowledge is expressed with the material or the building blocks of an object, such as food being made of food ingredients. A useful distinction between these three notions of part-whole: physical part of (sunroof - car), member of (musician - duet), and substance of (steel - boiler), is provided by WebChild. The importance of this commonsense dimension is shown by HasPartKB [7], which is an entire resource dedicated to part-whole relations.

Spatial. Spatial relations describe terms relating to or occupying space. This may entail indicating the usual location of a concept, as in the `location` property in wikidata or the `AtLocation` in ConceptNet. ConceptNet expresses locations for geographic entities, for example Boston is at location Massachusetts, as well as for things that can contain things: butter is at location refrigerator. Similar to the latter case, Wikidata includes an example that toasters are located in kitchens. A weaker spatial relation is one of spatial proximity in WebChild or ConceptNet, specifying that, e.g., bikes are located near roads. While Visual Genome does not explicitly have a spatial relation, concepts occurring in the same image region can be represented with the `LocatedNear` relation [35]. Example such statements include food being located near a plate or a table.

Creation. This dimension describes the process or the agent that brought something into existence. ConceptNet gives an example that a cake is created by the bake process, COMET has information that food is created from plants, while Wikidata states that rifle factories create shotguns. Table 2 reveals that no other source has creation information.

Utility. This dimension covers a notion of fitness or usefulness of objects for some purpose. ConceptNet's relation `UsedFor` expresses knowledge that 'the purpose of A is B', with an example of food being used for pleasure or nourishment. Wikidata has several similar relations: *use*, *used by*, and *uses*, which can express that platter is used for food presentation, or food is used by organisms. ConceptNet includes the notion of `CapableOf`, described as 'A is capable of B if A can typically do B', like food being capable of spoiling or "going rotten", or knives being capable of cutting. Another related notion is that of receiving an action: a button may receive the push action. While a button does not have the sole purpose of being pushed, it is capable of receiving that action, and by inference, it may respond to the action.

Desire or goal. This dimension covers knowledge about an agent's motivation, i.e., their desires or goals. An agent may want to have something or wish for something to happen. The agent typically has certain goals, aims, and/or plans, that may motivate or explain those desires. The relation `Desires` in ConceptNet may indicate, e.g., that a person desires regular access to food. Its negated version, `NotDesires` expresses that a person does not desire poisoned food. ATOMIC has two relations: `xWant` and `oWant`, to indicate the desires of an agent or other agents in a given situation. For instance, when people watch a movie together, they want to get some food. Regarding goals, ConceptNet includes the `MotivatedByGoal` and `ObstructedBy` relations to indicate the motivation and the constraint for a certain action. For instance, ConceptNet indicates that one's sleep is obstructed by noise, while COMET's extension of ConceptNet posits that people cook a meal because they are hungry.

Quality. Commonsense sources typically describe attributes of an agent or qualities related to an object. For example, ConceptNet and COMET include the relation `HasProperty`, to express knowledge like ice having the property cold and that food has property tasty. ATOMIC uses `xAttr` to indicate that, for example, the person that cooks food often has the attribute hungry or creative. WebChild and Wikidata both provide more specific qualities, such as taste, temperature, shape, or color. For instance, WebChild would specify the plant color as green. SenticNet describes the perceived "polarity" of an object, e.g., cold food is generally seen as negative.

Comparative. WebChild performs comparison of objects based on relative values for their attributes. Example comparative relations in WebChild are: healthier than (home cooking - fast food), faster than (car - bike), and larger than (lion - hyena). Notably, no other source describes comparative knowledge explicitly.[3]

---

[3]Here we exclude implicitly comparative knowledge, such as the inferred information that eating food makes one more satisfied from the triple: PersonX eats food - xReact - satisfied. We also include a separate category

Table 3
Examples for *food* for each of the 13 dimensions. When the subject is different from food, we state it explicitly, e.g., xWant: watch movie together – get some food.

Dimension	Example	Source
lexical	derivationaly related form: nutrient	WordNet
	etymologically related: fodder	ConceptNet
	derived term: foodie	ConceptNet
similarity	synonym: dish	ROGET
	said to be the same as: nutriment	Wikidata
	similar to: wholesome - organic	ConceptNet
distinctiveness	opposite of: non-food item	Wikidata
	distinct from: drink	ConceptNet
	different from: food safety - food security	Wikidata
taxonomic	hyponym: comfort food	WordNet
	hyponym: beverage	WordNet
	hypernym: substance	WordNet
	subclass of: disposable product	Wikidata
part-whole	things with food: minibar	ConceptNet
	is part of: life	COMET
	material used: food ingredient	Wikidata
spatial	is located at: pantry	ConceptNet
	is located at: a store	ConceptNet
	location: toaster - kitchen	Wikidata
	located near: plate	Visual Genome
	located near: table	Visual Genome
creator	is created by: cook	COMET
	is created by: plant	COMET
utility	use: eating	Wikidata
	used by: organism	Wikidata
	used for: pleasure	ConceptNet
	used for: sustain life	COMET
	used for: nourishment	ConceptNet
	capable of: cost money	ConceptNet
	capable of: go rotten	ConceptNet
	is capable of: taste good	COMET
goal/desire	xWant: watch movie together - get some food	ATOMIC
	desires: regular access to food	ConceptNet
	not desires: food poisoning	ConceptNet
	causes desire to: eat	ConceptNet
	xIntent: eats food - quit feeling hungry	ATOMIC
	motivated by: cook a meal	ConceptNet
	is motivated by: you be hungry	COMET
quality	xAttr: makes food - creative	ATOMIC
	has quality: shelf life	Wikidata
	has the property: tasty	COMET
comparative	healthier: home cooking - fast food	WebChild
temporal	has first subevent: cooking	ConceptNet
	starts with: open your mouth	COMET
	has effect: food allergy	Wikidata
	causes: you get full	COMET
	causes: indigestion	COMET
relational-other	related to: refrigerator	ConceptNet
	related to: cereal	ConceptNet
	field of work: food bank - food assistance	Wikidata
	main subject: cuisine - food product	Wikidata

Temporal. Most sources have notions of time that may support ordering by time and/or may capture relations that

---

for temporal representations that would include temporal comparison as one type of temporal information.

one thing is a prerequisite for another or one thing may have a particular effect. ConceptNet, for example, expresses that the first event of eating may be cooking, while the last one could be getting rid of the containers. COMET states that

---

eating starts with opening one's mouth. More strongly, the temporal relations often indicate relative ordering of two events, through relations of causation and effects, such as food potentially causing allergy or indigestion. Such causal knowledge is found in ATOMIC, ConceptNet, COMET, WebChild, and Wikidata.

Relational-other. Conceptual and context-related relationships are often underspecified. Increasingly resources are considering the setting or circumstances that surround an event or statement. ConceptNet has a single relation `HasContext` for this, while Wikidata has more concrete contextual relations, such as `field of this occupation`, `depicts`, and `health specialty`. This allows Wikidata to express that the main subject of a cuisine is a food product, and that the field of work of food banks is food assistance. Some resources have a kind of "catch-all" relationship that can be used very broadly. For example, most of the knowledge in ConceptNet belongs to a generic relation called `RelatedTo` that may be used to capture a relatively vague semantic connection between two concepts, such as food being related to refrigerator or cereal.

Our organization of existing relations into 13 dimensions provides a unified framework to reorganize and consolidate these sources. We also mention two nuances of our process. First, we placed the negative statements (marked with ¬ in Table 2) in the same dimension as the positive ones, as they cover the same knowledge type, despite having a different polarity and, arguably, purpose. Following a similar line of reasoning, we also placed inverse relations, such as used for and uses, in the same dimension. Second, we recognize that the underlying data may not always be clearly placed in only one of these dimensions. For instance, the relation `AtLocation`, which intuitively should belong to the spatial category, contains some statements that express part-whole knowledge.

5. Experiments

Seven of the sources covered in the previous section: ConceptNet, ATOMIC, Visual Genome, WordNet, Roget, Wikidata-CS, and FrameNet, have been integrated together in CSKG [35]. We start with CSKG and apply our dimension classification (section 4) to its sources, under an assumption that each of their edge relations can be mapped unambiguously to exactly one of the dimensions that fits it the best.[4] As a result, each edge in CSKG has dimension information stored in its `relation;dimension` column.[5] CSKG contains knowledge for 12 out of our 13 dimensions - the dimension *comparative* is not represented, as its only source, Web Child, is currently not part of CSKG. The resulting file is publicly available at: http://shorturl.at/msEY5.

This enrichment of the CSKG graph allows us to study the commonsense knowledge dimensions from multiple novel

perspectives. We investigate the following questions:

1. *Experiment 1:* How well is each dimension covered in the current sources? Here we compute the number of edges for each dimension across sources.
2. *Experiment 2:* Is knowledge redundant across sources? We use the dimensions to quantify overlap between sources with respect to individual edges.
3. *Experiment 3:* How do the edge dimensions compare to their lexical encoding? We compute clusters based on our dimensions and compare them to clusters computed with Transformer-based language models, like BERT [17] and RoBERTa [51].
4. *Experiment 4:* What is the impact of each dimension on downstream reasoning tasks? Each of the dimensions is used to select a subset of the available knowledge in CSKG. The selected knowledge is then used to pretrain a RoBERTa language model, which is applied to answer commonsense questions in a zero-shot manner.

In this section, we formulate and run suitable studies for each of the four questions, and reflect on the results.

5.1. Experiment 1: How well is each dimension covered in the current sources?

We use the CSKG graph enriched with edge dimensions to compute source coverage with respect to each dimension. The coverage of each source, formalized as a count of the number of edges per dimension, is presented in Table 4.[6]

We observe several trends in this Table. First, we recognize imbalance between the number of sources per dimension. Comparative knowledge and creation information are very rare and are described by only one or two sources, whereas taxonomic, temporal, and similarity knowledge are much more common and are captured by most sources. Second, some of the dimensions, like creation or part-whole, are represented with relatively few edges, whereas similarity and taxonomic knowledge generally have a much larger number of edges. The exception for the former is the large number of part-whole statements in WebChild, which is due to the fact that WebChild is automatically extracted, resulting in many duplicates and noisy information. Third, we see that some sources, like ConceptNet, FrameNet, and Wikidata-CS, aim for breadth and cover most dimensions. Others, like Roget and ATOMIC, have a narrow focus on specific dimensions: primarily desires/goals and temporal knowledge in ATOMIC, and only knowledge on similarity and distinctness in Roget. Yet, the narrow focus generally coincides with much depth, as both sources have many edges for the small set of dimensions that they cover. FrameNet, having a broad focus, has a small number of edges for each dimension due to its limited coverage of lexical units. Again here, WebChild is a notable outlier with a large number of automatically extracted statements for most dimensions. Finally, we observe

---

[4]As discussed before, this assumption might not always hold in practice. Future work should attempt to refine this mapping, e.g., by crowd-sourcing or by clustering algorithms.

[5]We leave out the relations prefixed with `/r/dbpedia` from Concept-Net, as these are being deprecated according to the official documentation: https://github.com/commonsense/conceptnet5/wiki/Relations.

[6]Python script: https://github.com/usc-isi-i2/cskg/blob/master/consolidation/compute_dimensions.py.

Table 4

Coverage of sources in terms of the knowledge dimensions. The numbers presented are in thousands.

Dimension	ATOMIC	ConceptNet	WebChild	Roget	Wikidata-CS	WordNet	FrameNet
lexical		704			0.5	207	14
similarity		255	343	1,023	1	152	0.4
distinctness		22		381	7	4	
taxonomic		244	783		73	89	23
part-whole		19	5,752		8	22	
spatial		28	660		0.5		
creation		0.3			0.2		
utility		69	2,843		2		1
desire/goal	244	20					
quality	143	9	6,510		1		11
comparative			813				
temporal	346	71	2,135		3		0.6
relational-other		1,969	291		6		0.7

Table 5

Overlap between various source pairs, based on the original relations (`Relations`) or the abstracted dimensions (`Dimensions`). Absolute overlap numbers are accompanied in brackets by the Jaccard percentage of the overlap against the union of all triples in the two sources.

Source pair	Relations	Dimensions
CN - RG	57,635 (1.23%)	73,992 (1.60%)
CN - WD	2,386 (0.07%)	2,623 (0.08%)
CN - WN	86,006 (2.14%)	97,946 (2.60%)
RG - WD	299 (0.02%)	333 (0.02%)
RG - WN	75,025 (3.55%)	75,025 (3.93%)
WD - WN	1,697 (0.19%)	1,704 (0.25%)

different ratios between 'strong' and 'weak' semantic relations across sources. Most of ConceptNet's knowledge falls under the generic *relational-other* category, whereas only a small portion of Wikidata-CS belongs to the same dimension. Most of Wikidata-CS is taxonomic knowledge.

5.2. Experiment 2: Is knowledge redundant across sources?

Our analysis so far reveals that most dimensions are covered by more than one source. This leads us to the next question: how often is a statement found in multiple sources?

Computing edge overlap between sources is conditioned on identity mapping between their nodes and relations. While CSKG provides such identity mappings between some of its nodes, this cannot be expected to be complete. We align the edges as follows. The **nodes** across sources are naively compared through their labels.[7] Regarding the **relations**, we benefit from CSKG's principle of normalizing the relations across sources to a shared set. With this procedure, a WordNet edge (`food.n.01, synonym, dish.n.01`) is modelled as (`food, /r/Synonym, dish`) in CSKG. As a dimension-based

enhancement, we abstract each relation further by mapping it to our dimensions, e.g., transforming (`food, /r/Synonym, dish`) to (`food, similarity, dish`). This dimension-based transformation allows for more flexible matching within a dimension, for instance, enabling similarity and synonymy statements to be compared for equivalence, since both (`food, /r/Synonym, dish`) and (`food, /r/SimilarTo, dish`) would be normalized to (`food, similarity, dish`).

We apply the relation-based and dimension-based variants to compute overlap between four sources: ConceptNet, Roget, Wikidata, and WordNet, in terms of each dimension. Here we do not consider ATOMIC or FrameNet, as their edges can be expected to have extremely low lexical overlap with the other sources. The overlap is computed as a Jaccard score between the number of shared triples between two sources and the union of their triples. The obtained scores are given in Table 5.[8] We observe that the overlap is generally low, yet, translating the original relations into dimensions constantly leads to an increase of the overlap for any of the source pairs. The highest relative overlap is observed between Roget and WordNet (3.93%), with ConceptNet-WordNet coming second (2.60%). The lowest overlap is obtained between the sources Roget and Wikidata (0.02%).

Next, we inspect the overlap between these sources for each dimension. Will the edges that correspond to more commonly found dimensions (e.g., part-whole, cf. Experiment 1) occur more often in multiple sources? We provide insight into this question in Table 6. Primarily, this Table reveals that there is very little edge overlap across sources. As hypothesized, most of the shared edges belong to dimensions that are common in many commonsense sources, describing `taxonomic`, `similarity`, and `part-whole` knowledge. The highest Jaccard score is obtained on the taxonomic knowledge between ConceptNet and WordNet, followed by similarity knowledge in ConceptNet-Roget and Roget-WordNet. Wikidata and ConceptNet share edges that belong to a num-

---

[7]If a node has more than one label, then we perform comparison based on the first one.

Table 6
Overlap distribution across dimensions. Absolute overlap numbers are accompanied in brackets by the Jaccard percentage of the overlap against the union of all triples in the two sources. '-' indicates that at least one of the sources does not use the dimension.

Sources	part-whole	taxonomic	lexical	distinctness	similarity	quality	utility	creation	temporal	rel-other
CN-RG	-	-	-	4,639	69,353	-	-	-	-	-
	-	-	-	(1.17)	(5.79)	-	-	-	-	-
CN-WD	68	1,888	20	266	102	0	14	0	1	264
	(0.25)	(0.62)	(0.00)	(1.00)	(0.04)	(0.00)	(0.02)	(0.00)	(0.00)	(0.01)
CN-WN	4,710	73,123	-	1,053	19,060	-	-	-	-	-
	(4.10)	(15.19)	-	(4.65)	(5.05)	-	-	-	-	-
RG-WD	-	-	-	206	127	-	-	-	-	-
	-	-	-	(0.05)	(0.01)	-	-	-	-	-
RG-WN	-	-	-	3,300	71,725	-	-	-	-	-
	-	-	-	(0.87)	(6.50)	-	-	-	-	-
WD-WN	82	1,533	-	63	26	-	-	-	-	-
	(0.07)	(0.39)	-	(0.62)	(0.02)	-	-	-	-	-

ber of other dimensions, including `distinctness`, `similarity`, and `rel-other`.

The sparse overlap in Tables 5-6 is amplified by our lexical method of computing overlap, as the same or similar nodes may have slightly different labels. Both the low overlap and the relatively weak comparison method strongly motivate future work on node resolution of commonsense KGs. Node resolution of commonsense KGs is non-trivial, as nodes are intended to represent various aspects of meaning: words, phrases, concepts, frames, events, and sentences. We discuss further the challenges and opportunities relating to node resolution of commonsense KGs in Section 7.

5.3. Experiment 3: How do the edge dimensions compare to their lexical encoding?

Next, we investigate how the information captured by our dimensions relates to the encoding of edges by state-of-the-art Transformer-based language models, like BERT or RoBERTa.[9] For this purpose, we cluster the knowledge in CSKG according to our 13 dimensions, resulting in 13 disjoint clusters. We also compute clusters based on language models in an unsupervised manner as follows. Each of the edges is lexicalized into a natural language sentence by relation-specific templates. Each sentence is then encoded with a Transformer model, either BERT-large or RoBERTa-large, into a single 1,024-dimensional embedding. These embeddings are finally clustered with the k-Means [29] algorithm into $k = 13$ disjoint clusters.

The two approaches for computing clusters, based on our dimensions and based on Transformer embeddings, can now be compared in terms of their agreement. We use the adjusted rand index (ARI) metric to measure agreement.[10] The ARI score is 0.226 for BERT and 0.235 for RoBERTa. These

scores signal low agreement between the dimension-based and the unsupervised clustering, which is expected given that the dimension-based clustering entirely depends on the relation, while the unsupervised clustering considers the entire triple. We also observe that the ARI score of RoBERTa is slightly higher, which might indicate that the relation has higher impact on the embedding in RoBERTa than in BERT.
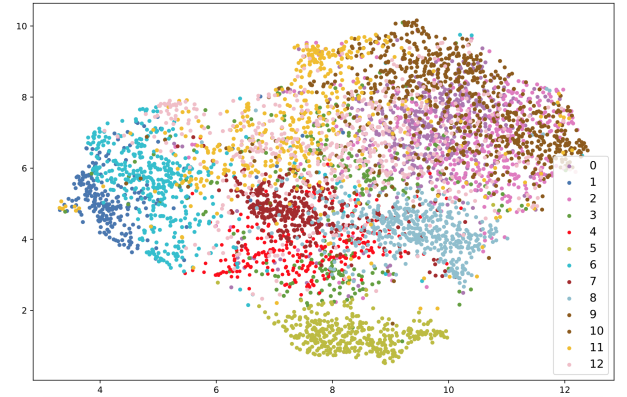

Figure 1: UMAP clusters of RoBERTa.

In order to understand the information encoded by RoBERTa, we use UMAP to visualize its k-means clusters for a random sample of 5,000 edges (Figure 1). Curiously, certain clusters are clearly delineated, while others are not. For instance, cluster 5 has little overlap with the other clusters. Looking into the contents of this cluster, we observe that it is largely dominated by distinctness information: 92% (360 out of 390) of its edges belong to this dimension, mostly expressed through the `/r/Antonym` relation. Clusters 4, 7, and 8 are largely dominated by similarity, while clusters 1 and 6 are largely split between temporal (46%) and desire/goal (36%) edges. At the same time, we observe a lot of over-

Table 7

10 pairs with the highest Jaccard scores between dimension-based and RoBERTa-based clusters.

RoBERTa cluster	Dimension cluster	Jaccard
5	distinctness	0.916
8	similarity	0.452
6	temporal	0.322
7	similarity	0.295
6	desire/goal	0.283
1	desire/goal	0.258
4	similarity	0.210
0	lexical	0.205
1	temporal	0.202
12	relational-other	0.182

laps between the clusters 0, 2, 9, 10, 11, and 12. These clusters are dominated by lexical and relational-other edges - e.g., around half of all edges in clusters 0 and 9 belong to the category relational-other. The node frequency distributions reveal that cluster 1 describes positive emotions, as its most frequent node is /c/en/happy; nodes in cluster 5 are often numbers, like rg:en_twenty-eighth; and the nodes in cluster 9 describe concepts from natural sciences, the most connected node being /c/en/zoology.
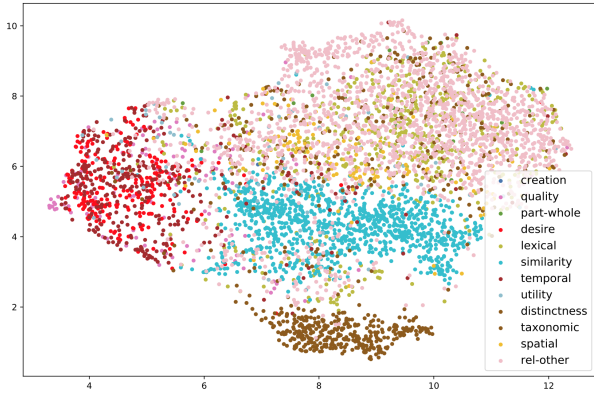
Figure 2: UMAP clusters according to our dimensions.

In Figure 2, we visualize the same set of edges, only this time we color each according to their dimension. In accordance with the relatively low rand index score, we observe that the clusters are mostly not well-distinguished from one another. We look for correspondences between the RoBERTa clusters in Figure 1 and the dimension-based clusters in Figure 2, by computing Jaccard score between the edges that constitute each pair of their clusters. The 10 cluster pairs with the highest Jaccard scores are shown in Table 7. We observe the highest correspondence for distinctness with cluster 5 (Jaccard score of 0.92) and similarity with cluster 8 (score 0.45). The overlapping clusters for desire and temporal knowledge both map relatively strongly to clusters 1 and 6, which appear nearby in the RoBERTa clustering as well.
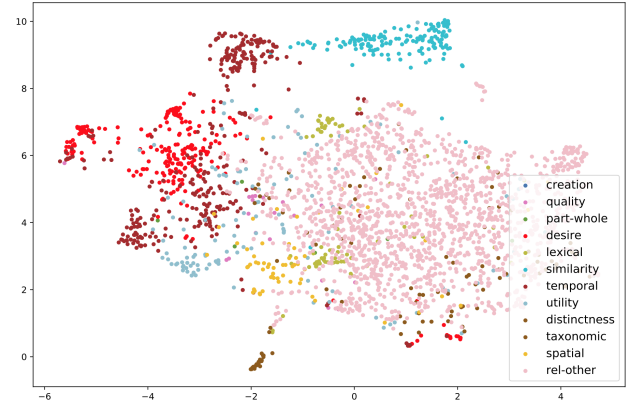
Table 8

Top-3 highest-scored dimensions for each of the automatically-computed clusters. Bold-faced results indicate the top score for each dimension.

0	**lexical (0.205)**, rel-other (0.121), taxonomic (0.060)
1	**desire (0.258)**, temporal (0.202), quality (0.087)
2	lexical (0.133), rel-other (0.122), taxonomic (0.075)
3	**spatial (0.119)**, lexical (0.061), quality (0.053)
4	similarity (0.21), quality (0.015), lexical (0.008)
5	**distinctness (0.916)**, lexical (0.018), taxonomic (0.003)
6	**temporal (0.322)**, desire (0.283), quality, (0.054)
7	similarity (0.295), quality (0.009), taxonomic (0.005)
8	**similarity (0.452)**, lexical (0.004), taxonomic (0.003)
9	rel-other (0.143), lexical (0.085), **taxonomic (0.081)**
10	rel-other (0.169), taxonomic (0.015), quality (0.007)
11	rel-other (0.11), **quality (0.068)**, taxonomic (0.066)
12	**rel-other (0.183)**, taxonomic (0.059), spatial (0.053)

We also observe relatively high score for the cluster pairs 4-similarity, 7-similarity, and 0-lexical, all of which confirm our prior analysis of Figure 1.

We show the top-3 highest-scored dimensions for each of the automatically-computed clusters in Table 8. Here, we observe that some dimensions, like creation, utility, and part-whole do not fall within the top-3 dimensions for any of the RoBERTa clusters. This is likely due to the lower number of edges with these dimensions, as well as their dispersion across many RoBERTa clusters.

Figure 3: UMAP clusters for two selected nodes: /c/en/food and /c/en/eat.

To investigate further, we select the two CSKG nodes we considered earlier in Table 4: /c/en/food and /c/en/eat, and visualize their edges according to their dimension. The total set of 2,661 edges largely belongs to the dimension *relational-other* (1,553), followed by *temporal* (319 edges), and *desire/goal* (228 edges), whereas no edge belongs to the *creation* dimension. Thus, most of the well-specified edges about nutritional concepts express either temporal information about the process, or knowledge about desires/goals relating to nutrition. Within a single dimension, most *relational-other* edges are expressed with RelatedTo (1,488 out of 1,553).

Temporal knowledge is split into multiple relations, primarily `HasLastSubevent` (113 edges), `HasPrerequisite` (69), and `HasSubevent` (60). *Desire/goal* divided into `at:xWant` (78 edges), `MotivatedByGoal` (47), and `at:xIntent` (47 edges). Besides the two seed nodes (`/c/en/food` and `/c/en/eat`), the frequency distribution of nodes in a cluster reveals other prominent nodes. Naturally, the spatial dimension includes `/c/e/plate` (with an edge degree of 3), the temporal cluster includes `/c/en/diminish_own_hunger` (degree of 4), and the distinctness cluster has 7 edges for `/c/en/drink`.

In summary, experiment 3 reveals a relatively low agreement between the dimensions and the language model-based clustering. This indicates that the nodes of an edge, which are not considered by the dimensions, are important for the language models. Meanwhile, we observe individual correspondences between the language models and our dimensions, most notably on the dimensions of similarity and distinctness. In addition, both the dimensions and the language models naturally cluster nodes from different sources together.

5.4. Experiment 4: What is the impact of each dimension on downstream reasoning tasks?

Experiment 3 revealed overlaps between the information captured by our dimensions and that captured by language models. In our fourth experiment, we experiment with enhancing language models with knowledge belonging to individual dimensions, in order to examine the effect of different dimensions of knowledge on commonsense reasoning tasks.

We adopt the method proposed by Ma et al. [53] to pretrain state-of-the-art language models and conduct zero-shot evaluation on two commonsense question answering tasks. According to this method, we first transform ConceptNet, WordNet, Wikidata, ATOMIC, and Visual Genome into synthetic QA sets. We use templates to map each triple into a QA pair and apply random sampling with heuristic-based filtering to collect two distractors for every QA pair. We group synthetic QA pairs based on their dimension, resulting in 12 dimension-based QA buckets in total. Within each dimension, the QA data is split into a training and development set. For ATOMIC, we adopt its original split to partition data. For the other sources, we select 95% of the data as a training set and the remaining 5% as a development set, following the original paper. We note that Ma et al. [53] only selected 14 relations from ConceptNet, WordNet, Wikidata, and Visual Genome, whereas we include all relations except `RelatedTo`. The statistics for the synthetic QA sets are shown in Table 9. We can see that the knowledge distribution across dimensions is fairly skewed, with creation having very few questions, while taxonomic and temporal knowledge being very well populated. Our experiments in this section will reveal whether the amount of available knowledge affects downstream task performance.

We pretrain the RoBERTa-large [51] model on each of the dimensions using the corresponding synthetic QA set. We use RoBERTa with a marginal ranking objective, as this is the best combination according to [53]. We use the same set of hyper-parameters as in [53], except for the creation

Table 9
Statistics of the number of QA pairs for each dimension.

Dimension	Train	Dev
part-whole	87,765	4,620
taxonomic	340,609	17,927
lexical	107,861	5,677
distinctness	20,286	1,068
similarity	166,575	8,768
quality	116,593	12,492
utility	63,862	3,362
creation	304	17
temporal	312,628	31,587
relational-other	242,759	12,777
spatial	21,726	1,144
desire/goal	194,906	20,912

dimension. Specifically, we train our models for 1 epoch using a learning rate of $1e-5$, a batch size of 32, and 1.0 as a margin. For the creation dimension, since the number of samples is much smaller, we train the model for 5 epochs while keeping the other hyper-parameters fixed. We evaluate our models on two tasks: the CommonSenseQA (CSQA) task [83], in which the model is asked to choose the correct answer for a given question out of five options, and the SocialIQA (SIQA) task [73], in which the model chooses the correct answer from three options based on a question and a brief context.

The results from our experiments are shown in Table 10. Overall, we see that the zero-shot pretraining on any dimension allows the model to outperform the no-knowledge baseline. However, the variance of the improvement across dimensions is relatively large, revealing that certain dimensions are more relevant for the two downstream tasks than others. For example, although the training set size of *lexical* dimension exceeds 107K, its performance gain on both tasks is limited. This is not surprising, since the language model can be expected to already have learned most of the lexical knowledge during its pretraining on unstructured text corpora. While the *quality* dimension has a similar training set size as the *lexical* dimension, the model benefits from it by a large margin on both tasks: 20.7 and 12.7 absolute points, respectively. This finding suggests that learning additional knowledge on object qualities is novel and useful, as it has not been fully acquired from unstructured data.

We note that reasoning on downstream tasks benefits especially from the knowledge dimensions that align with the task question types. For example, each question in SIQA corresponds to an ATOMIC relation, and requires knowledge primarily about the order of events, personal attributes, and agent desires. Consequently, pretraining on *quality*, *temporal*, and *desire/goal* knowledge provides the model with the largest gain on the SIQA task. The improvements of the *temporal* dimension are even higher than training on the entire set of questions, suggesting that certain knowledge dimensions that are not related to social reasoning may even lead to a decline in model performance. For the CSQA task,

Table 10
Zero-shot evaluation results on two commonsense reasoning tasks. We run every experiment 3 times with different seeds, and report mean accuracy with a 95% confidence interval.

Dimension	CSQA	SIQA
Baseline	45.0	47.3
+part-whole	63.0(±1.4)	52.6(±1.9)
+taxonomic	62.6(±1.4)	52.2(±1.6)
+lexical	49.9(±2.9)	49.0(±0.4)
+distinctness	57.2(±0.5)	50.2(±1.5)
+similarity	61.4(±0.8)	53.5(±0.6)
+quality	65.7(±0.5)	60.0(±0.7)
+utility	**67.4(±1.0)**	54.8(±0.7)
+creation	49.9(±1.1)	47.8(±0.2)
+temporal	67.3(±0.3)	**62.6(±0.9)**
+relational-other	58.2(±1.7)	51.3(±1.7)
+spatial	63.3(±0.2)	53.1(±0.3)
+desire/goal	65.0(±1.8)	60.0(±0.6)
+all	66.2(±1.4)	61.0(±0.7)

Table 11
Zero-shot evaluation results of RoBERTa on the synthetic QA sets.

Dimension	Dev accuracy
part-whole	67.5
taxonomic	57.0
lexical	90.1
distinctness	77.3
similarity	65.6
quality	45.5
utility	67.9
creation	82.4
temporal	47.2
relational-other	37.6
spatial	56.9
desire/goal	48.0

since it is derived from the broad set of knowledge dimensions covered in ConceptNet, we expect that many (if not all) of these dimensions would help performance. Accordingly, we observe large gains with many of the knowledge dimensions (+15%), whereas training on the utility dimension yields the best performance (+22.4%), even slightly better than training on the entire set of questions.

To improve our understanding of the impact of every dimension of knowledge on different questions, we further break down the performance of the models per question type. For CSQA, we classify the questions based on the Concept-Net relations between the correct answer and the question concept, as the model needs to reason over such relations in order to be able to answer the question. For SIQA, since the questions are initially generated using a set of templates based on ATOMIC relations, we try to reverse-engineer the process by manually defining a mapping from the question format to the ATOMIC relations. Using this method, we are able to successfully classify more than 99% of questions from SIQA's development set. For every question type, we compute the average accuracy over 3 seeds for the models trained on each dimension. The results on CSQA and SIQA are shown in the Figures 4 and 5, respectively.[11]

For some question types in CSQA, (one of) the largest improvements is achieved by training on the corresponding knowledge dimension. For example, *temporal* knowledge for questions of type `Causes`, and *desire/goal* for `Desires` questions. However, in other cases, the accuracy improvement by the corresponding knowledge dimension is much lower than other dimensions, for example, *distinctness* knowledge is less useful for `Antonym` questions compared to *utility* knowledge. This could be interpreted as a signal that the knowledge might not be clearly separated between the dimensions. For SIQA, the results show that the corresponding knowledge dimension is clearly impactful for most question types:

desire/goal on `xWant`, `xIntent`, *quality* on `xAttr` and *temporal* on `xNeed`, `oReact`, `oEffect`. This is especially visible for `xIntent` and `xNeed`, where very little gain is observed for most knowledge dimensions except for the corresponding dimension, suggesting that aligning the questions to their knowledge dimensions is important for building effective models. We note that a similar finding on the alignment between knowledge and the task has been reported in the original paper [53]; yet, the dimensions allow us to validate this claim more precisely.

Finally, to verify our hypothesis that certain dimensions of knowledge are already learned by the state-of-the-art language models to a large extent, while others are not, we directly evaluate the baseline LM on the synthetic QA sets for each dimension. The results are shown in Table 11. As expected, even without any training, the model already achieves a very high accuracy (over 90%) on the *lexical* dimension, thus it could not receive much training signal from this dimension. Meanwhile, the LM accuracy on *quality*, *temporal*, and *desire/goal* dimensions is significantly lower. This is mostly because questions from ATOMIC take a large portion in these dimension. As reported in [53], the questions from ATOMIC are more challenging than those created from other knowledge resources. We note that the accuracy for *relational-other* is the lowest among all dimensions. We hypothesize that this is because this dimension is noisier than others and the knowledge in this dimension is less likely to be found in the unstructured text. We leave further investigation of this result for future research.

In summary, we observe that certain dimensions of knowledge are very beneficial and novel for language models, allowing them to improve their performance on downstream reasoning tasks. Other dimensions, like lexical knowledge, are almost entirely redundant, as the language models have already acquired this knowledge during their initial training. The exact contribution for each dimension depends on the knowledge required by the task at hand. We discuss the implications of the obtained results further in Section 7.

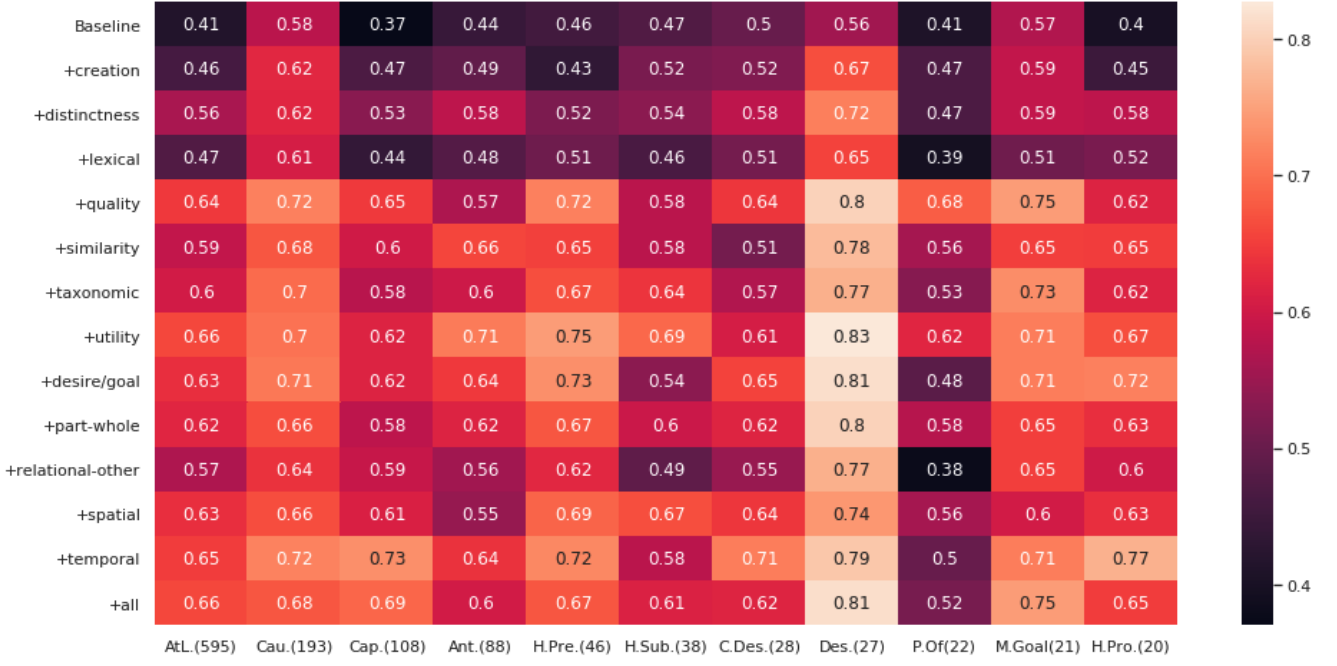[11]For CSQA, we omit question types with less than 20 questions.

Figure 4: Accuracy for each question type in CSQA, where **AtL.** means AtLocation, **Cau.** means Causes, **Cap.** means CapabelOf, **Ant.** means Antonym, **H.Pre.** means HasPrerequisite, **H.Sub.** means HasSubevent, **C.Des.** means CauseDesires, **Des.** means Desires, **P.Of** means ParfOf, **M.Goal** means MotivatedByGoal, **H.Pro** means HasProperty. The numbers in parentheses indicate how many questions fall into the category.
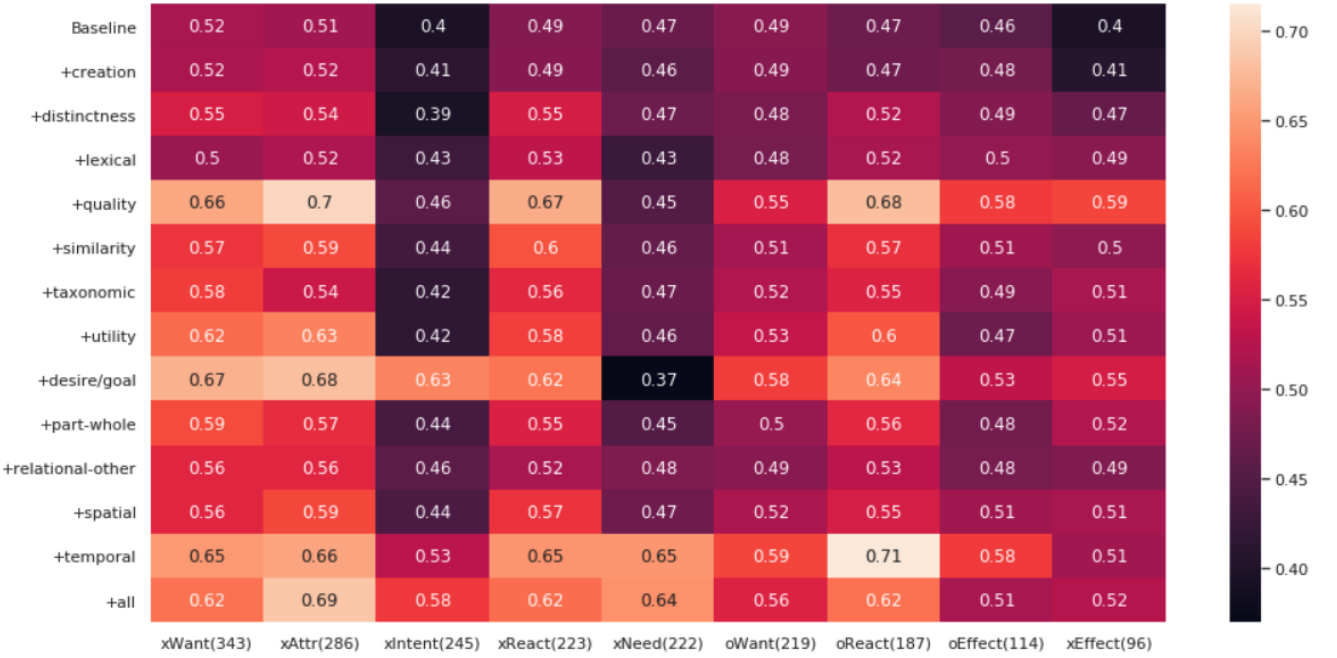

Figure 5: Accuracy for each question type in SIQA. The numbers in parentheses indicate how many questions fall into that category.

6. Related Work

6.1. Epistemic Foundations

World knowledge is what people learn and generalize from physical and social experiences, distilling mental representations out of the most significant aspects of everyday life.[12] In these terms, commonsense knowledge can be con-

---

[12] According to the classic argument of the mind-body problem, it is inherently impossible to characterize how generalization occurs, due to an *explanatory gap* [44].

ceived as the partition of world knowledge that is *commonly shared* by most people. This definition, however, has intrinsic limitations: in fact, the scale and the diversity of physical and social experiences, as well as the ecological, context-dependent nature of what constitutes 'common' and 'uncommon' [25], make it hard to formulate any abstract criterion of what should fall under commonsense knowledge. From Aristotle's theory of categories [1] to Brentano's empirical psychology [11], deriving knowledge dimensions from empirical observations, as opposed to from *abstract criteria* [31], is a fundamental epistemic approach that contributed to the birth of Cognitive Science as a discipline [57], besides serving as reference framework for our current investigation. In this article, in fact, we neither propose nor adopt any *a priori* principle to define what should be included in a commonsense knowledge graph; rather, we analyze multiple knowledge graphs and, supported by empirical methods and experimental validations, elicit their most salient conceptual structures.

In the history of general knowledge base systems, the difficulty of characterizing common sense has been a driver, rather than an obstacle. For instance, Cyc [43], the most monumental effort to construct an axiomatic theory of commonsense knowledge, has been actively growing for almost forty years. At present, the Cyc knowledge base comprises around 1.5 million general concepts, and 25 million rules and assertions about these concepts. Different domain-specific extensions of Cyc exist, funded by industry and government programs: considering the key role that commonsense knowledge can play in enhancing AI systems for private and public enterprises, Cyc's strategy of steering the general knowledge base development in the direction of domain use cases can represent a sustainable business model for other stakeholders in the field. Since our focus is on computational resources of commonsense knowledge, it is beyond the scope of this paper to discuss seminal work on axiomatization of common sense, such as Pat Hayes' *naive physics* [30], Gordon and Hobbs' *commonsense psychology* [26], and Ken Forbus' work on *qualitative commonsense reasoning* [21]. Similarly, although we recognize the important role that semantic discourse theory plays in describing commonsense utterances based on expressive logical criteria [2, 42], it is beyond the scope of the current contribution to consider whether the knowledge resources we illustrated can be used as a basis to study the formal structures of dialog and, in general, sociolinguistic interaction, grounded on common sense.

Existing commonsense knowledge graphs make implicit categorizations of knowledge, by defining a tractable set of relations which can be traced to some types proposed in cognitive research. For instance, WebChild's [84] part-of relations resemble the partonomic-metonimic relations in cognitive science literature (e.g., see [14, 89]), while ConceptNet [81] defines 34 relations, where the relation IsA can be often approximated with taxonomic knowledge. In its first version [50], ConceptNet defined 20 relations grouped into 8 categories: K-lines, Things, Agents, Events, Spatial, Causal, Functional, and Affective. Zhang et al. [93] extrapolate the

types in the Conceptual Semantic Theory with those in ConceptNet 1.0 and propose the following six categories: property, object, eventuality, spatial, quantity, and others. According to [54], four types of contextual knowledge are essential for humans to interpret or frame in text: intratextual, intertextual, extratextual, and circumtextual knowledge. These generic knowledge types, which are orthogonal to the declarative/procedural distinction, are relevant for most natural language understanding tasks. In [36], we analyzed these types for the task of entity linking; when it comes to commonsense question answering, they might provide a guide for extending the coverage of the knowledge, when combined with specific theories/axioms, such as those defined by the resources in Section 3.2 or in Cyc.

Beyond the structural differences, commonsense knowledge graphs share the same foundational elements. Commonsense knowledge is generally split into declarative and procedural, where the former is contained in unconditional assertions, and the latter requires conditional assertions[13]: we can state, for instance, that windows are *typically* made of glass (declarative), and assert that if a large rock is thrown against a window, the glass typically breaks (procedural). As the use of the adverb *typically* suggests, commonsense knowledge rules out exceptions from the context of interpretation: for instance, bulletproof glass doesn't break when hit by a rock.

There is a consensus that commonsense knowledge describes the world in terms of objects with their properties and relations. Yet, there is no consensus in terms of how this knowledge should be formalized and used for reasoning. Moore [59] argues that reasoning over existentially quantified prepositions, negations, or by cases, all require an approach based on formal logic. State-of-the-art techniques for commonsense reasoning [47, 52, 53] are largely based on statistical (language) models, complemented with knowledge represented in a textual form [72, 81].

Statistical models are largely based on co-occurrence patterns and primarily capture contextual similarity of concepts, e.g., based on lexical co-occurrence and structured patterns [70]. Recent work has investigated the extent to which purely statistical models, such as the Transformer models, can perform reliable and generalizable inference. The work by Ettinger [20] shows that BERT struggles with complex inference, role-based event prediction, and the contextual impacts of negation. Bhagavatula et al. [5] reveal that pretrained language models perform relatively well on commonsense inference, except on tasks that require numeric and spatial understanding. The logical commonsense probes in RICA [94] show that LMs perform similar to random guessing in the zero-shot setting, they are heavily impacted by statistical biases, and are not robust to linguistic perturbations. This is in line with the findings of Elazar et al. [19], which indicate that the good performance of LMs mostly comes from fine-tuning on the task data. Recognizing that commonsense reasoning largely relies on pragmatic understanding [27], Jeretič

---

[13]The distinction between these types of assertions was formalized in a seminal work by Gentzen [24].

et al. [37] demonstrate that BERT exhibits partial ability for pragmatic reasoning on certain scalar implicatures and presuppositions.

It has been shown that language models have acquired certain commonsense knowledge during their pre-training on large textual data [64, 16]. This inspired follow-up work that queried/prompted language models directly, treating them as knowledge bases [79, 77]. Alternatively, language models can be pretrained with structured knowledge and applied to downstream reasoning tasks in a zero-shot manner [53, 4]. Considering the large capacity of the pre-trained language models and the goal of generalizing to unseen tasks [53], lightweight alternatives to fine-tuning have been proposed recently. One approach is to preserve the knowledge acquired during pretraining, but complement it by updating a small amount of additional parameters [49, 46]. Alternatively, one could keep the entire model intact and tune the model input for optimal knowledge elicitation [39, 77]. Ultimately, it is unclear whether aspects like negation and case-based reasoning can be fully solved by language models, or require inclusion of structured knowledge and formal logic, as suggested by Moore [59].

6.2. Consolidation efforts

In this paper, we analyzed individual knowledge graphs through suitable semantic dimensions, with the goal of providing insights on how alignment and consolidation of commonsense knowledge resources can be guided and, eventually, achieved. A natural extension of our work would be to evaluate ongoing efforts that adopt alternative methods of consolidation: accordingly, Framester [22], BabelNet [61], CSKG [35], and Predicate Matrix [41] constitute some of the most mature projects in this space. In Framester, several resources like WordNet, VerbNet, FrameNet, and BabelNet, are aligned using an OWL schema based on *Description and Situations* and *Semiotics* ontology design patterns.[14] CSKG, which we leverage in this paper, is also based on a schema, but it doesn't rely on traditional RDF/OWL semantics: in fact, CSKG is a hyper-relational graph represented in a tabular format, designed to preserve individual knowledge structures of resources like ConceptNet, WebChild, Visual Genome, etc., exploit direct mappings when available, derive indirect mappings when possible (e.g., while ConceptNet and Visual Genome do not have direct connections, they both have mappings to WordNet), and infer links through statistical algorithms. BabelNet is a multilingual lexicalized semantic network based on automatically linking Wikipedia with WordNet, and expanded by using additional information from resources like FrameNet and VerbNet. Finally, the Predicate Matrix exploits Word Sense Disambiguation algorithms to generate semi-automatic mappings between FrameNet, VerbNet, PropBank, WordNet, and the Events and Situations Ontology [76].[15]

---

[14] These patterns can be accessed at http://ontologydesignpatterns.org/wiki/Main_Page

[15] https://github.com/newsreader/eso-and-ceo

Ji et al. [38] survey knowledge graphs in terms of their representation principles, acquisition methods, and applications. This work considers commonsense reasoning as a key emerging application area for knowledge graphs. A review of the state-of-the-art commonsense reasoning is provided by Storks et al. [82], with a strong focus on benchmarks and reasoning techniques. Recently, several tutorials on commonsense knowledge and reasoning have been held, with focus on consolidation [33], neural model reasoning [74], neuro-symbolic integration [32], acquisition [85, 68], and broader applications [85].

7. Discussion and Roadmap

7.1. Summary of Findings

Commonsense knowledge sources use different levels of semantics, come in a variety of forms, and strive to capture diverse notions of common sense. After surveying 20 commonsense knowledge sources, we proposed that their relations can be grouped into 13 dimensions, namely: lexical, similarity, distinctness, part-whole, spatial, creation, utility, desire/goal, quality, comparative, temporal, and relational-other. Most relations can be unambiguously mapped to one of the dimensions. We apply our dimensions to reorganize the knowledge in the CSKG graph [35]. Following our devised mapping of relations to dimensions, we add an additional column (relation;dimension) in CSKG, indicating the dimension of each edge. This allows us to make use of the consolidation of seven existing sources done by CSKG, and complement it with the dimensions in order to perform more abstract analysis of its knowledge types. We designed and ran four experiments to analyze commonsense knowledge in CSKG through the lenses of these 13 dimensions.

In experiment 1, we investigated the coverage of the 13 dimensions in current sources. Some dimensions, like part-whole and similarity, are a subject of interest in most sources. Others, like comparative knowledge and knowledge on desires/goals, are rarely captured. Yet, the depth of knowledge on the less commonly represented relations is still high, as illustrated by the 244 thousand desire/goal edges in ATOMIC. Here we also observed that the breadth of focus varies notably across sources, as some (e.g., ConceptNet and Wikidata-CS) cover a wide range of relations, while others (e.g., ATOMIC or WordNet) have a narrower focus.

Experiment 2 posed the question of whether individual knowledge statements are redundant across sources. Our experiments with four sources indicated, with few exceptions, that only a tiny portion of all edges were shared between a pair of sources. This experiment points to a two-fold motivation for node resolution over commonsense sources. On one hand, node resolution is needed to increase the quality of computing overlap beyond the current lexical comparison of nodes. On the other hand, as the sources have generally complementary goals, it is likely that even with a more semantic computation the overlap will remain low. Node resolution is, thus, essential to consolidate different views of a node (concept) into a single representation.

In experiment 3, we cluster all edges in CSKG according to their dimension, and compare these clusters with $k = 13$ clusters based on a language model encoding of each edge. We noted that the overall agreement between the dimensions and the language model-based clustering is relatively low, indicating that language models pay much attention to the edge nodes. However, individual correspondences were noted. Similarity and distinctness quite clearly dominated some of the RoBERTa-based clusters, while other clusters were consistently split between the dimensions of desire/goal and temporal knowledge. Interestingly, the clusters inferred from the RoBERTa embeddings often clustered nodes from different sources into a single cluster.

Finally, in experiment 4 we investigated the impact of the dimensions on a downstream reasoning task of commonsense question answering. We adopted a recent idea [53] for pretraining language models with knowledge graphs. The best-scoring model in this paper, RoBERTa-large with marginal ranking loss, was provided with knowledge from one of our dimensions at a time, and evaluated in a zero-shot manner on two benchmarks testing broad (CommonsenseQA) and social (SocialIQA) commonsense reasoning. The experiments showed that social commonsense reasoning clearly benefits from temporal, quality, and desire/goal knowledge, whereas the CommonsenseQA benchmark benefits from broad knowledge from all dimensions. Certain dimensions, such as lexical knowledge, were relatively uninformative, as it can be expected that such knowledge has been already acquired by the language models at their initial training stage. While the extent of knowledge plays a role, adding more knowledge is not always beneficial, as the task performance depends on the alignment between the dimensions and the task. This motivates further work on automatic alignment between the task questions and the our dimensions. This also motivates future work that attempts to evaluate the value of additional content additions related to particular dimensions aimed at improving certain kinds of tasks.

7.2. Outlook

The goal of consolidating and applying commonsense knowledge is an ambitious one, as witnessed by decades of research on this topic. Our 13 dimensions are an effort to reorganize existing commonsense knowledge through unification of its knowledge types. We see this as a necessary, but not sufficient, step towards a modern and comprehensive commonsense resource. The dimensions could facilitate, or complement, several other aspects of this pursuit:

1. Node resolution While potentially controversial, the consolidation of commonsense knowledge relations into dimensions/knowledge types is achievable with careful manual effort. This is largely due to the relatively small number of relations in most current sources. Resolving nodes across sources is another key aspect of this consolidation, strongly motivated by experiment 2 of this paper. Sources capture complementary knowledge, whose combination is prevented by the lack of mappings between their nodes. As nodes are currently intended to represent various aspects of meaning:

words, phrases, concepts, frames, events, sentences — their consolidation is not obvious. Moreover, the number of unique nodes in most sources is on the order of many thousands or even millions [35], preventing it from being solvable by mere manual effort. Node resolution could be framed as a 'static' disambiguation/clustering task, where each statement for a node is to be classified into one of its meanings, similar to [15]. Here, the set of meanings can be either present (e.g., WordNet synsets) or dynamically inferred from the data. An alternative, 'dynamic' approach is defer the node resolution to task time and perform it implicitly as a task of retrieving evidence from a knowledge source [45]. Another option is a combination of the static and the dynamic approaches.

2. Coverage and boundaries At present, it is difficult to estimate the completeness of commonsense knowledge sources. With the relations organized into dimensions of knowledge, we gain insight into the volume of knowledge that falls within each of the dimensions. An ideal node resolution would take us one step further, allowing us to detect gaps, i.e., understand which relevant facts are not represented by any of the sources. If nodes are resolved to an ontology like WordNet, one could leverage its taxonomy to infer new information. For instance, ConceptNet is at present unable to infer that if barbecues are held in outdoor places, they could be, by extension, be held in a park or someone's patio. In addition, a more semantic resource would allow us to define constraints over the knowledge and detect anomalies and contradictory knowledge, which can be argued to define the boundaries of the knowledge that can be obtained. It is reasonable that such boundaries exist, as commonsense knowledge is characterized by commonness of its concepts and restricted set of relations [34]. Further, organizing by dimensions also allows us to describe strengths (and/or weaknesses) of resources. For example, a resource that has many partonomic relationships might be the first resource to consider using if a task requires part-whole reasoning.

3. Generalizable downstream reasoning As current large-scale commonsense sources are primarily text-based, they are lexicalized prior to their combination with language models, losing much of their structure. As this lack of structure prevents us from understanding their coverage and gaps, we are unable to measure their potential for downstream reasoning as a function of the available knowledge. It remains unknown to which extent a more complete source, organized around dimensions of commonsense knowledge, would be able to contribute to improving performance. Experiment 4 showed that there is correspondence between knowledge dimensions and question answering tasks, motivating automatic alignment between the two. Moreover, a comprehensive semantic source may inspire new neuro-symbolic reasoning methods, with potentially enhanced generalizability and explainability, opening the door for reliable commonsense services to be made available in the future.

4. Evaluation and knowledge gaps Experiment 4 showed that the potential of different dimensions for reasoning varies greatly and is largely dependent on the evaluation data. This

finding is in line with [53]. The fact that certain dimensions consistently contribute little can be an indicator of gaps in current evaluation. Namely, dimensions like *distinctness* and *spatial* which currently contribute little or not at all are likely to be underrepresented in current evaluations. These gaps should ideally be addressed in the future by new benchmarks that will represent these missing dimensions. We note that our set of dimensions is based on the relations found in current popular commonsense sources. Hence, in this paper, we make an assumption that the knowledge types in these sources suffice (or at least previously sufficed) and can express the desired knowledge. The diversity of knowledge expressed by the *relational-other* dimension, as pointed out also in [34], might be an indicator for additional, latent dimensions hidden behind the vagueness of this dimension.

8. Conclusions

At present, commonsense knowledge is dispersed across a variety of sources with different foci, strengths, and weaknesses. The complementary knowledge covered by these sources motivates efforts to consolidate them into a common representation. In this paper, we pursued the goal of organizing commonsense relations into a shared set of knowledge dimensions in a bottom-up fashion. Starting from a survey and analysis of the relations found in existing sources, we grouped them into 13 dimensions: lexical, similarity, distinctness, part-whole, spatial, creation, utility, desire/goal, quality, comparative, temporal, and relational-other. As each relation in these sources can be mapped to a dimension, we applied our method to abstract the relations in an existing consolidated resource: the Commonsense Knowledge Graph. This allowed us to empirically study the impact of these dimensions. First, we observed that some dimensions are included more often than others, potentially pointing to gaps in the knowledge covered in existing resources. Second, we measured sparse overlap of facts expressed with each dimension across sources, which motivates future work on graph integration through (automated) node resolution. Third, comparing the dimension-based clustering to language model-based unsupervised edge clustering resulted in low overall agreement, though in some cases, the unsupervised clusters were dominated by one or two dimensions. This showed that some of the dimensions represent a stronger signal for language modeling than others. Fourth, we measured the impact of each dimension on a downstream question answering reasoning task, by adapting a state-of-the-art method of pre-training language models with knowledge graphs. Here, we observed that the impact differs greatly by dimension, depending largely on the alignment between the task and the knowledge dimension, as well as on the novelty of knowledge captured by a dimension. While this is in accordance with the findings of the original method [53], the dimension-driven experiments of this paper enabled this hypothesis to be investigated much more precisely, revealing the direct impact of each knowledge dimension rather than entire knowledge sources.

Our experiments inspired a four-step roadmap towards creation and utilization of a comprehensive dimension-centered resource. (1) Node resolution methods should be introduced and applied to unify the resources further. (2) Such an integration would allow us to better understand and improve the coverage/gaps and boundaries of these sources. (3) A large-scale, public semantic graph of commonsense knowledge may inspire novel neuro-symbolic methods, potentially allowing for better generalization and explainability. (4) The impact of a dimension is an indicator of the coverage of that dimension in current evaluation benchmarks; under-represented dimensions are evaluation gaps that may need to be filled by introducing new benchmarks. And, vice-versa, additional knowledge dimensions might be hidden behind the generic *relational-other* dimension.

Acknowledgement

References

[1] Aristotle, Jones, R., Ross, W., 2012. The Metaphysics. CreateSpace Independent Publishing Platform. URL: https://books.google.com/books?id=-D3JtAEACAAJ.

[2] Asher, N., Morreau, M., 1990. Commonsense entailment: A modal theory of nonmonotonic reasoning, in: European Workshop on Logics in Artificial Intelligence, Springer. pp. 1–30.

[3] Baker, C.F., Fillmore, C.J., Lowe, J.B., 1998. The berkeley framenet project, in: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1, pp. 86–90.

[4] Banerjee, P., Baral, C., 2020. Self-supervised knowledge triplet learning for zero-shot question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online. pp. 151–162. URL: https://www.aclweb.org/anthology/2020.emnlp-main.11, doi:10.18653/v1/2020.emnlp-main.11.

[5] Bhagavatula, C., Bras, R.L., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., tau Yih, W., Choi, Y., 2020. Abductive commonsense reasoning, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=Byg1v1HKDB.

[6] Bhakthavatsalam, S., Anastasiades, C., Clark, P., 2020a. Genericskb: A knowledge base of generic statements. arXiv preprint arXiv:2005.00660 .

[7] Bhakthavatsalam, S., Richardson, K., Tandon, N., Clark, P., 2020b. Do dogs have whiskers? a new knowledge base of haspart relations. arXiv preprint arXiv:2006.07510 .

[8] Bisk, Y., Zellers, R., LeBras, R., Gao, J., Choi, Y., 2020. Piqa: Reasoning about physical commonsense in natural language., in: AAAI, pp. 7432–7439.

[9] Boratko, M., Li, X.L., Das, R., O'Gorman, T.J., Le, D., McCallum, A., 2020. Protoqa: A question answering dataset for prototypical common-sense reasoning, in: EMNLP.

[10] Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Çelikyilmaz, A., Choi, Y., 2019. Comet: Commonsense transformers for automatic knowledge graph construction, in: ACL.

[11] Brentano, F., 2014. Psychology from an empirical standpoint. Routledge.

[12] Cambria, E., Li, Y., Xing, F.Z., Poria, S., Kwok, K., 2020. Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment

analysis. Proceedings of the 29th ACM International Conference on Information & Knowledge Management .

[13] Cambria, E., Song, Y., Wang, H., Howard, N., 2012. Semantic multi-dimensional scaling for open-domain sentiment analysis. IEEE intelligent systems 29, 44–51.

[14] Casati, R., Varzi, A.C., et al., 1999. Parts and places: The structures of spatial representation. Mit Press.

[15] Chen, J., Liu, J., 2011. Combining conceptnet and wordnet for word sense disambiguation, in: Proceedings of 5th International Joint Conference on Natural Language Processing, pp. 686–694.

[16] Davison, J., Feldman, J., Rush, A., 2019. Commonsense knowledge mining from pretrained models, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China. pp. 1173–1178. URL: https://www.aclweb.org/anthology/D19-1109, doi:10.18653/v1/D19-1109.

[17] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 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, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota. pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423, doi:10.18653/v1/N19-1423.

[18] Dodge, E.K., Hong, J., Stickles, E., 2015. Metanet: Deep semantic automatic metaphor analysis, in: Proceedings of the Third Workshop on Metaphor in NLP, pp. 40–49.

[19] Elazar, Y., Zhang, H., Goldberg, Y., Roth, D., 2021. Back to square one: Bias detection, training and commonsense disentanglement in the winograd schema. arXiv preprint arXiv:2104.08161 .

[20] Ettinger, A., 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the Association for Computational Linguistics 8, 34–48.

[21] Forbus, K.D., 1997. Qualitative reasoning.

[22] Gangemi, A., Alam, M., Asprino, L., Presutti, V., Recupero, D.R., 2016. Framester: A wide coverage linguistic linked data hub, in: European Knowledge Acquisition Workshop, Springer. pp. 239–254.

[23] Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L., 2002. Sweetening ontologies with dolce, in: International Conference on Knowledge Engineering and Knowledge Management, Springer. pp. 166–181.

[24] Gentzen, G., 1934. Investigations into logical deduction. translation printed in m. szabo the collected papers of gerhard gentzen.

[25] Gibson, E.J., Pick, A.D., et al., 2000. An ecological approach to perceptual learning and development. Oxford University Press, USA.

[26] Gordon, A.S., Hobbs, J.R., 2017. A formal theory of commonsense psychology: How people think people think. Cambridge University Press.

[27] Grice, H.P., 1975. Logic and conversation, in: Speech acts. Brill, pp. 41–58.

[28] Guha, R.V., Brickley, D., Macbeth, S., 2016. Schema. org: evolution of structured data on the web. Communications of the ACM 59, 44–51.

[29] Hartigan, J.A., Wong, M.A., 1979. Algorithm as 136: A k-means clustering algorithm. Journal of the royal statistical society. series c (applied statistics) 28, 100–108.

[30] Hayes, P.J., 1990. The naive physics manifesto, in: Boden, M.A. (Ed.), The Philosophy of Artificial Intelligence. Oxford University Press. Oxford readings in philosophy, pp. 171–205.

[31] Hicks, G.D., 1904. Idealism and the problem of knowledge and existence, in: Proceedings of the Aristotelian Society, JSTOR. pp. 136–178.

[32] Ilievski, F., Bosselut, A., Razniewski, S., Kejriwal, M., 2021a. Commonsense knowledge acquisition and representation, in: AAAI Tutorials.

[33] Ilievski, F., Szekely, P., Kejriwal, M., 2020a. Commonsense knowledge graphs, in: The International Semantic Web Conference (ISWC)

[34] Ilievski, F., Szekely, P., Schwabe, D., 2020b. Commonsense knowledge in wikidata. Wikidata workshop .

[35] Ilievski, F., Szekely, P.A., Zhang, B., 2021b. CSKG: the commonsense knowledge graph, in: Verborgh, R., Hose, K., Paulheim, H., Champin, P., Maleshkova, M., Corcho, Ó., Ristoski, P., Alam, M. (Eds.), The Semantic Web - 18th International Conference, ESWC 2021, Virtual Event, June 6-10, 2021, Proceedings, Springer. pp. 680–696. URL: https://doi.org/10.1007/978-3-030-77385-4_41, doi:10.1007/978-3-030-77385-4_41.

[36] Ilievski, F., Vossen, P., Van Erp, M., 2017. Hunger for contextual knowledge and a road map to intelligent entity linking, in: International Conference on Language, Data and Knowledge, Springer, Cham. pp. 143–149.

[37] Jeretic, P., Warstadt, A., Bhooshan, S., Williams, A., 2020. Are natural language inference models IMPPRESsive? Learning IM-Plicature and PRESupposition, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 8690–8705. URL: https://www.aclweb.org/anthology/2020.acl-main.768, doi:10.18653/v1/2020.acl-main.768.

[38] Ji, S., Pan, S., Cambria, E., Marttinen, P., Philip, S.Y., 2021. A survey on knowledge graphs: Representation, acquisition, and applications. IEEE Transactions on Neural Networks and Learning Systems .

[39] Jiang, Z., Xu, F.F., Araki, J., Neubig, G., 2020. How can we know what language models know? Transactions of the Association for Computational Linguistics 8, 423–438. URL: https://www.aclweb.org/anthology/2020.tacl-1.28, doi:10.1162/tacl_a_00324.

[40] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al., 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision 123, 32–73.

[41] de Lacalle, M.L., Laparra, E., Aldabe, I., Rigau, G., 2016. A multilingual predicate matrix, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 2662–2668.

[42] Lascarides, A., Asher, N., 1993. Temporal interpretation, discourse relations and commonsense entailment. Linguistics and philosophy 16, 437–493.

[43] Lenat, D.B., 1995. Cyc: A large-scale investment in knowledge infrastructure. Communications of the ACM 38, 33–38.

[44] Levine, J., 1983. Materialism and qualia: The explanatory gap. Pacific philosophical quarterly 64, 354–361.

[45] Lewis, P.S.H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (Eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

[46] Li, X.L., Liang, P., 2021. Prefix-tuning: Optimizing continuous prompts for generation. arXiv:2101.00190.

[47] Lin, B.Y., Chen, X., Chen, J., Ren, X., 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), Association for Computational Linguistics, Hong Kong, China. pp. 2829–2839. URL: https://www.aclweb.org/anthology/D19-1282, doi:10.18653/v1/D19-1282.

[48] Lin, B.Y., Lee, S., Khanna, R., Ren, X., 2020a. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online. pp. 6862–6868. URL: https://www.aclweb.org/anthology/2020.

emnlp-main.557, doi:10.18653/v1/2020.emnlp-main.557.

[49] Lin, Z., Madotto, A., Fung, P., 2020b. Exploring versatile generative language model via parameter-efficient transfer learning, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online. pp. 441–459. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.41, doi:10.18653/v1/2020.findings-emnlp.41.

[50] Liu, H., Singh, P., 2004. Conceptnet—a practical commonsense reasoning tool-kit. BT technology journal 22, 211–226.

[51] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692.

[52] Ma, K., Francis, J., Lu, Q., Nyberg, E., Oltramari, A., 2019. Towards generalizable neuro-symbolic systems for commonsense question answering, in: Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing, Association for Computational Linguistics, Hong Kong, China. pp. 22–32. URL: https://www.aclweb.org/anthology/D19-6003, doi:10.18653/v1/D19-6003.

[53] Ma, K., Ilievski, F., Francis, J., Bisk, Y., Nyberg, E., Oltramari, A., 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering, in: AAAI.

[54] MacLachlan, G.L., Reid, I., 1994. Framing and interpretation. Melbourne University Press.

[55] McCarthy, J., et al., 1960. Programs with common sense. RLE and MIT computation center.

[56] Miller, G.A., 1998. WordNet: An electronic lexical database. MIT press.

[57] Miller, G.A., 2003. The cognitive revolution: a historical perspective. Trends in cognitive sciences 7, 141–144.

[58] van Miltenburg, E., 2016. Stereotyping and bias in the flickr30k dataset, in: Edlund, J., Heylen, D., Paggio, P. (Eds.), Proceedings of Multimodal Corpora: Computer vision and language processing (MMC 2016), pp. 1–4. URL: http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-MCC-2016-proceedings.pdf.

[59] Moore, R.C., 1982. The role of logic in knowledge representation and commonsense reasoning. SRI International. Artificial Intelligence Center.

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

[61] Navigli, R., Ponzetto, S.P., 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193, 217–250.

[62] Niles, I., Pease, A., 2001. Towards a standard upper ontology, in: Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001, pp. 2–9.

[63] Onyeka, E., Varde, A., Anu, V., Tandon, N., Daramola, O., 2020. Using commonsense knowledge and text mining for implicit requirements localization. 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI) , 935–940.

[64] Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., Miller, A., 2019. Language models as knowledge bases?, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China. pp. 2463–2473. URL: https://www.aclweb.org/anthology/D19-1250, doi:10.18653/v1/D19-1250.

[65] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S., 2016. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. International Journal of Computer Vision , 1–20.

[66] Puri, M., Varde, A., Dong, B., 2018. Pragmatics and semantics to connect specific local laws with public reactions. 2018 IEEE International Conference on Big Data (Big Data) , 5433–5435.

[67] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language models are unsupervised multitask learners. OpenAI Blog 1, 9.

[68] Razniewski, S., Tandon, N., Varde, A.S., 2021. Information to wisdom: Commonsense knowledge extraction and compilation, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 1143–1146.

[69] Roget, P.M., 2020. Roget's Thesaurus. Good Press.

[70] Rohde, D.L., Gonnerman, L.M., Plaut, D.C., 2006. An improved model of semantic similarity based on lexical co-occurrence. Communications of the ACM 8, 116.

[71] Romero, J., Razniewski, S., Pal, K., Z. Pan, J., Sakhadeo, A., Weikum, G., 2019. Commonsense properties from query logs and question answering forums, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 1411–1420.

[72] Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N.A., Choi, Y., 2019a. Atomic: An atlas of machine commonsense for if-then reasoning, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3027–3035.

[73] Sap, M., Rashkin, H., Chen, D., Le Bras, R., Choi, Y., 2019b. Social IQa: Commonsense reasoning about social interactions, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China. pp. 4463–4473. URL: https://www.aclweb.org/anthology/D19-1454, doi:10.18653/v1/D19-1454.

[74] Sap, M., Shwartz, V., Bosselut, A., Choi, Y., Roth, D., 2020. Commonsense reasoning for natural language processing, in: ACL.

[75] Schuler, K.K., 2005. Verbnet: A broad-coverage, comprehensive verb lexicon .

[76] Segers, R., Vossen, P., Rospocher, M., Serafini, L., Laparra, E., Rigau, G., 2015. Eso: A frame based ontology for events and implied situations. Proceedings of MAPLEX 2015.

[77] Shin, T., Razeghi, Y., IV, R.L.L., Wallace, E., Singh, S., 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts, in: Webber, B., Cohn, T., He, Y., Liu, Y. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics. pp. 4222–4235. URL: https://doi.org/10.18653/v1/2020.emnlp-main.346, doi:10.18653/v1/2020.emnlp-main.346.

[78] Shwartz, V., West, P., Le Bras, R., Bhagavatula, C., Choi, Y., 2020a. Unsupervised commonsense question answering with self-talk, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online. pp. 4615–4629. URL: https://www.aclweb.org/anthology/2020.emnlp-main.373, doi:10.18653/v1/2020.emnlp-main.373.

[79] Shwartz, V., West, P., Le Bras, R., Bhagavatula, C., Choi, Y., 2020b. Unsupervised commonsense question answering with self-talk, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online. pp. 4615–4629. URL: https://www.aclweb.org/anthology/2020.emnlp-main.373, doi:10.18653/v1/2020.emnlp-main.373.

[80] Singh, P., Lin, T., Mueller, E.T., Lim, G., Perkins, T., Zhu, W.L., 2002. Open mind common sense: Knowledge acquisition from the general public, in: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", Springer. pp. 1223–1237.

[81] Speer, R., Chin, J., Havasi, C., 2017. Conceptnet 5.5: An open multilingual graph of general knowledge, in: Thirty-First AAAI Conference on Artificial Intelligence.

[82] Storks, S., Gao, Q., Chai, J., 2019. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. ArXiv abs/1904.01172.

[83] Talmor, A., Herzig, J., Lourie, N., Berant, J., 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, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota. pp. 4149–4158. URL: https://www.aclweb.org/anthology/N19-1421, doi:10.18653/v1/N19-1421.

[84] Tandon, N., De Melo, G., Weikum, G., 2017. Webchild 2.0: Fine-grained commonsense knowledge distillation, in: Proceedings of ACL 2017, System Demonstrations, pp. 115–120.

[85] Tandon, N., Varde, A.S., de Melo, G., 2018. Commonsense knowledge in machine intelligence. ACM SIGMOD Record 46, 49–52.

[86] Tanon, T.P., Weikum, G., Suchanek, F.M., 2020. Yago 4: A reasonable knowledge base. The Semantic Web 12123, 583 – 596.

[87] Vrandečić, D., Krötzsch, M., 2014. Wikidata: a free collaborative knowledgebase. Communications of the ACM 57, 78–85.

[88] Williams, B., Lieberman, H., Winston, P.H., 2017. Understanding stories with large-scale common sense., in: COMMONSENSE.

[89] Winston, M.E., Chaffin, R., Herrmann, D., 1987. A taxonomy of part-whole relations. Cognitive science 11, 417–444.

[90] Wu, W., Li, H., Wang, H., Zhu, K.Q., 2012. Probase: A probabilistic taxonomy for text understanding, in: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp. 481–492.

[91] Yang, W., Wang, X., Farhadi, A., Gupta, A., Mottaghi, R., 2018. Visual semantic navigation using scene priors. arXiv preprint arXiv:1810.06543 .

[92] Zellers, R., Bisk, Y., Farhadi, A., Choi, Y., 2019. From recognition to cognition: Visual commonsense reasoning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6720–6731.

[93] Zhang, H., Zhao, X., Song, Y., 2020. WinoWhy: A deep diagnosis of essential commonsense knowledge for answering Winograd schema challenge, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 5736–5745. URL: https://www.aclweb.org/anthology/2020.acl-main.508, doi:10.18653/v1/2020.acl-main.508.

[94] Zhou, P., Khanna, R., Lee, S., Lin, B.Y., Ho, D., Pujara, J., Ren, X., 2020. Rica: Evaluating robust inference capabilities based on commonsense axioms. arXiv preprint arXiv:2005.00782 .