Chapter 6
ConceptNet 5: A Large Semantic Network for Relational Knowledge

Robyn Speer and Catherine Havasi

Abstract ConceptNet is a knowledge representation project, providing a large semantic graph that describes general human knowledge and how it is expressed in natural language. Here we present the latest iteration, ConceptNet 5, with a focus on its fundamental design decisions and ways to interoperate with it.

6.1 Introduction

The wisdom of crowds can be found all over the Web. Some of the most significant recent advances in collecting the world's knowledge appear in resources such as Wikipedia and Wiktionary, which are written for people by large numbers of people, yet converge on a structure that can be made understandable by computers. Meanwhile, "games with a purpose" collect large quantities of specific knowledge while simply providing entertainment in return. Both are knowledge sources that can provide a wealth of information to computers about how people use and understand language, as long as it can be compiled into a useful and scalable representation.

ConceptNet is a project that creates such a representation of crowd-sourced knowledge, providing a large semantic graph that describes general human knowledge and how it is expressed in natural language. The scope of ConceptNet includes words and common phrases in any written human language. It provides a large set of background knowledge that a computer application working with natural language text should know.

These words and phrases are related through an open domain of predicates, such as *IsA* or *UsedFor*, describing not just how words are related by their lexical

R. Speer (✉) · C. Havasi (✉)
MIT Media Lab, 20 Ames St., Cambridge, MA 02142, USA
e-mail: rspeer@arborelia.net; havasi@media.mit.edu

I. Gurevych and J. Kim (eds.), *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, DOI 10.1007/978-3-642-35085-6_6,

Table 6.1 The most common interlingual relations in ConceptNet, with example sentence frames in English and their number of collected edges

Relation	# edges	Sentence pattern
IsA	7,956,303	*NP* is a kind of *NP*.
PartOf	536,648	*NP* is part of *NP*.
AtLocation	535,278	Somewhere *NP* can be is *NP*.
RelatedTo	319,471	*NP* is related to *NP*.
HasProperty	303,921	*NP* is *AP*.
UsedFor	254,563	*NP* is used for *VP*.
DerivedFrom	242,853	*TERM* is derived from *TERM*.
Causes	233,727	The effect of *VP* is *NP*|*VP*.
CapableOf	167,405	*NP* can *VP*.
MotivatedByGoal	173,111	You would *VP* because you want *VP*.
HasSubevent	154,214	One of the things you do when you *VP* is *NP*|*VP*.
Desires	95,779	*NP* wants to *VP*.
HasPrerequisite	69,474	*NP*|*VP* requires *NP*|*VP*.
HasA	56,691	*NP* has *NP*.
CausesDesire	51,338	*NP* makes you want to *VP*.
MadeOf	43,278	*NP* is made of *NP*.
DefinedAs	39,406	*NP* is defined as *NP*.
HasFirstSubevent	35,242	The first thing you do when you *VP* is *NP*|*VP*.
ReceivesAction	24,609	*NP* can be *VP*.
LocatedNear	12,679	You are likely to find *NP* near *NP*.
SimilarTo	11,635	*NP* is like *NP*.
SymbolOf	11,302	*NP* represents *NP*.
HasLastSubevent	8,689	The last thing you do when you *VP* is *NP*|*VP*.
CreatedBy	1,979	You make *NP* by *VP*.

definitions, but also how they are related through common knowledge. We will refer to these as *relations*. The most common ones appear in Table 6.1.

For example, ConceptNet's knowledge about "jazz" includes not just the properties that define it, such as *IsA*(jazz, genre of music); it also includes incidental facts such as

- *AtLocation*(jazz, new orleans)
- *UsedFor*(saxophone, jazz), and
- *Plays percussion in*(jazz drummer, jazz).

A cluster of related concepts and the ConceptNet relations that connect them is visualized in Fig. 6.1.

ConceptNet originated as a representation for the knowledge collected by the Open Mind Common Sense project [21], which uses a long-running interactive Web site to collect new statements from visitors to the site, and asks them target questions about statements it thinks may be true. Later releases included knowledge from similar websites in other languages, such as Portuguese and Dutch, and collaborations with online word games that automatically collect general knowledge, yielding further knowledge in English, Japanese, and Chinese.

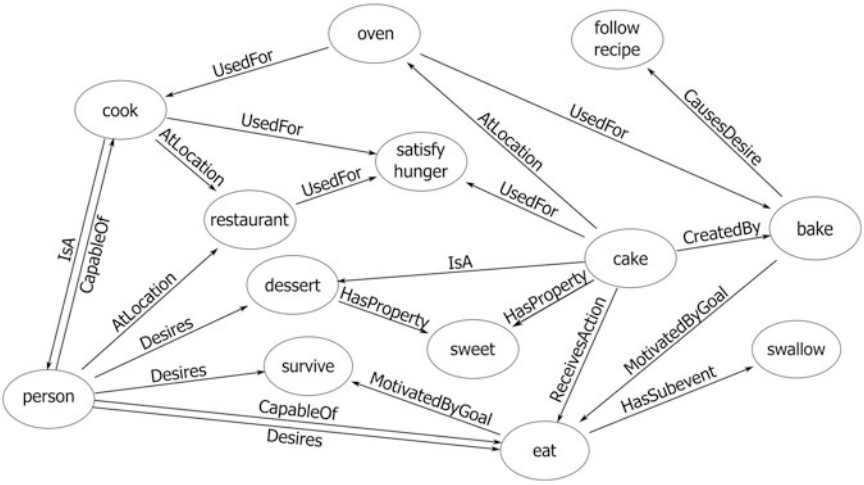

Fig. 6.1 A high-level view of the knowledge ConceptNet has about a cluster of related concepts

ConceptNet gives a foundation of real-world knowledge to a variety of AI projects and applications. Previous versions of ConceptNet [11] have been used, for example, to build a system for analyzing the emotional content of text [6], to create a dialog system for improving software specifications [14], to recognize activities of daily living [24], to visualize topics and trends in a corpus of unstructured text [23], and to create public information displays by reading text about people and projects from a knowledge base [12].

There are similar ongoing projects that collect crowd-sourced knowledge from similar sources. BabelNet [18] and MENTA [7], for example, create a large, structured, multilingual taxonomy from a combination of Wikipedia's structured knowledge and WordNet. WikiNet [17] also mines structured knowledge from Wikipedia, while a project by Blanco et al. [4] creates a ConceptNet-like representation whose input primarily comes from unstructured machine reading. ConceptNet's niche is defined by its representational decisions, which are particularly suited for some kinds of text understanding applications:

- Its concepts are connected to, and defined by, natural language words and phrases that can also be found in free text.
- It includes not just definitions and lexical relationships, but also the "common sense" associations that ordinary people make among these concepts. Its sources range in formality from dictionaries to online games.
- It puts more emphasis on collecting information about common words than about named entities (which is why, for example, it collects more from Wiktionary than it does from Wikipedia).
- The concepts are not limited to a single language; they can be from any written language.

- It integrates knowledge from sources with varying levels of granularity and varying registers of formality, and makes them available through a common representation.

ConceptNet aims to contain both specific facts and the messy, inconsistent world of *common sense knowledge*. To truly understand concepts that appear in natural language text, it is important to recognize the informal relations between these concepts that are part of everyday knowledge, which are often under-represented in other lexical resources. WordNet, for example, can tell you that a dog is a type of carnivore, but not that it is a type of pet. It can tell you that a fork is an eating utensil, but has no link between *fork* and *eat* to tell you that a fork is used for eating.

Adding common sense knowledge creates many new questions. Can we say that "a fork is used for eating" if a fork is used for other things besides eating, and other things are used for eating? Should we make sure to distinguish the eating utensil from the branching of a path? Is the statement still true in cultures that typically use chopsticks instead of forks? We can try to collect representations that answer these questions, while pragmatically accepting that much of the content of a common sense knowledge base will leave them unresolved.

6.1.1 Motivation for ConceptNet 5

In comparison to previous versions of ConceptNet, the new goals of ConceptNet 5 include:

- Incorporating knowledge from other crowd-sourced resources with their own communities and editing processes, particularly data mined from Wiktionary and Wikipedia.
- Adding links to other resources such as DBPedia [2], Freebase [5], and WordNet [10].
- Supporting machine-reading tools such as ReVerb [9], which extracts relational knowledge from Web pages.
- Finding translations between concepts represented in different natural languages.

ConceptNet 5 is intended to grow freely and absorb knowledge from many sources, with contributions from many different projects. We aim to allow different projects to contribute data that can easily be merged into ConceptNet 5 without the difficulty of aligning large databases.

Combining all these knowledge sources in a useful way requires processes for normalizing and aligning their different representations, while avoiding information loss. It also requires a system for comparing the reliability of the collected knowledge, as such knowledge can come from a variety of processes, sometimes involving unreliable sources (such as players of online games) and sometimes involving unreliable processes (parsers and transformations between representations).

In a sense, while ConceptNet 4 and earlier versions collected facts, ConceptNet 5 also at a higher level collects *sources* of facts. This greatly expands its domain, makes it interoperable with many other public knowledge resources, and makes it applicable to a wider variety of text-understanding applications.

6.2 Knowledge in ConceptNet 5

ConceptNet expresses *concepts*, which are words and phrases that can be extracted from natural language text; we call them "concepts" instead of terms to account for the fact that they can be more or less specific than a typical term. ConceptNet also contains *assertions* of the ways that these concepts relate to each other. These assertions can come from a wide variety of sources that create *justifications* for them. The current sources of knowledge in ConceptNet 5 are:

- The Open Mind Common Sense website (http://openmind.media.mit.edu), which collects common-sense knowledge mostly in English, but has more recently supported other languages.
- Sister projects to OMCS in Portuguese [1] and Dutch [8].
- The multilingual data, including translations between assertions, collected by the GlobalMind project, a spin-off of OMCS.
- "Games with a purpose" that collect common knowledge, including Verbosity [26] in English, *nadya.jp* in Japanese, and the "pet game" [16] on the popular Taiwanese bulletin board PTT, collecting Chinese knowledge in traditional script.
- A new process that scans the English Wiktionary (a Wikimedia project at en.wiktionary.org that defines words in many languages in English). In addition to extracting structured knowledge such as synonyms and translations, it also extracts some slightly-unstructured knowledge. For example, it extracts additional translations from the English-language glosses of words in other languages. The process is similar to that of UKPL [27] but targets ConceptNet's representation.
- WordNet 3.0 [10], including cross-references to its RDF definition at http://semanticweb.cs.vu.nl/lod/wn30/ [25].
- The semantic connections between Wikipedia articles represented in DBPedia [2], with cross-references to the corresponding DBPedia resources. DBPedia contains a number of collections, in different languages, representing relationships with different levels of specificity. So far we use only the English collection, and only use links that translate to our standard relations "IsA", "PartOf", and "AtLocation".
- Relational statements mined from Wikipedia's free text using ReVerb [9], run through a filter we designed to keep only the statements that are going to be most useful to represent in ConceptNet. We discarded statements whose ReVerb scores were too low, and those that contained uninformative terms such as "this".

Adding knowledge from other free projects such as WordNet does more than just increase the coverage of ConceptNet; it also allows us to align entries in ConceptNet with those in WordNet, and refer to those alignments without having to derive them again. This is an important aspect of the Linked Data movement: different projects collect data in different forms, but it is best when there is a clear way to map from one to the other. When the data is linked, ConceptNet enhances the power of WordNet and vice versa.

Adding data from Wiktionary was key in unifying the data collected in many different languages. In ConceptNet 4, each language was a separate connected component; now all the languages of ConceptNet are highly interlinked.

ConceptNet 5 is growing as we find new sources and new ways to integrate their knowledge. As of April 2012, it contains 12.5 million edges, representing about 8.7 million assertions connecting 3.9 million concepts. 2.78 million of the concepts appear in more than one edge. Its most represented language is English, where 11.5 million of the edges contain at least one English concept. The next most represented languages are Chinese (900,000 edges), Portuguese (228,000 edges), Japanese (130,000 edges), French (106,000 edges), Russian (93,700 edges), Spanish (92,400 edges), Dutch (90,000 edges), German (86,500 edges), and Korean (71,400 edges). The well-represented languages largely represent languages for which multilingual collaborations with Open Mind Common Sense exist, with an extra boost for languages that are well-represented in Wiktionary.

Additional sources that may be added include the plan-oriented knowledge in Honda's Open Mind Indoor Common Sense [13], connections to knowledge in Freebase [5], ontological connections to SUMO and MILO [19], and new processes that scan well-structured Wiktionaries in other target languages, such as Japanese and German.

6.2.1 Representation

ConceptNet 5 is conceptually represented as a hypergraph. Its assertions can be seen as edges that connect its nodes, which are concepts (words and phrases), via relations that are also nodes. These assertions, however, can be *justified* by other assertions, knowledge sources, or processes. The predicates that label them can be one of a set of interlingual relations, such as *IsA* or *UsedFor*, or they can be automatically-extracted relations that are specific to a language, such as *is_known_for*, or underspecified prepositional relations such as *is_on*.

The values of the predicates – referred to hereafter as the *relation* of each assertion – are represented using concept nodes as well. The structure of edges surrounding two assertions appears in Fig. 6.2. The most common interlingual relations we identify in ConceptNet appear in Table 6.1.

One way to represent a hypergraph is to reify all edges as nodes, with lower-level relationships such as "x is the first argument of y" becoming the new edges. We experimented with representations of reified hypergraphs, but found that the result

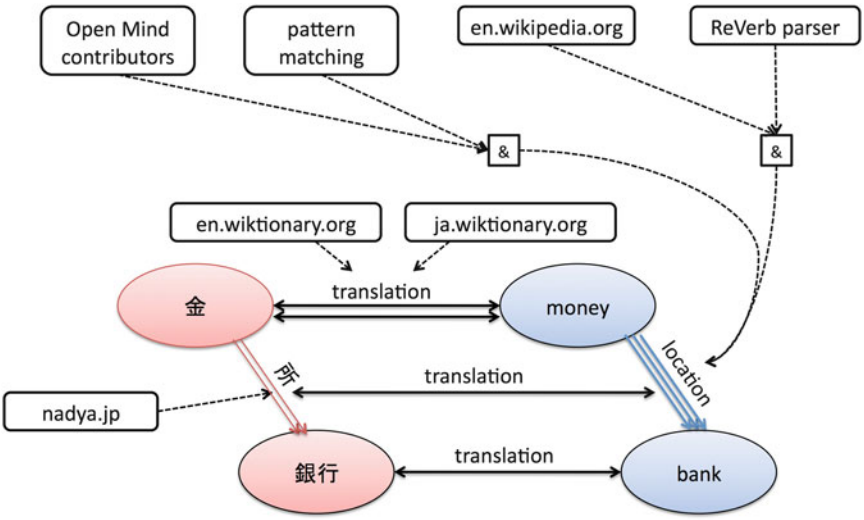

Fig. 6.2 An example of two assertions in ConceptNet 5, and the edges they involve. *Rounded rectangles* and *dotted edges* represent knowledge sources; *solid edges* are grouped together into assertions

was exceptionally difficult to query as the database grew. Asking simple questions such as "What are the parts of a car?" in a hypergraph is a complex, multi-step query, and we found no mature database system that could perform all the queries we needed efficiently.

Instead, we store almost all of the relevant information about an edge as properties on that edge. Each assertion is still reified with a unique ID, but that ID is only referred to within the assertion or in higher-level assertions about that assertion, such as translations.

In particular, an edge in ConceptNet 5 is an *instance* of an assertion, as learned from some knowledge source. The same assertion might be represented by a large bundle of edges, when we learn it in many different ways; these all have the same assertion ID, along with algorithmically-generated unique edge IDs that we can use to deduplicate data later.

A hypergraph can be represented in a standard graph format such as RDF, but only by reifying all of its edges. It is straightforward to export an RDF version of ConceptNet 5 that conveys the same information, but the overhead created by reifying everything would make it a poor choice for a native representation.

6.2.2 Assertion Scores

The sources that justify each assertion form a structure that can be seen as a disjunction of conjunctions. Each edge – that is, each instance of an assertion – indicates a combination of sources that produced that edge, while the bundle of

edges making up an assertion represents the disjunction of all those conjunctions. Examples of these structures appear in Fig. 6.2.

Each conjunction comes with a positive or negative *score*, a weight that it assigns to that edge, with more complex conjunctions having an inherently lower weight. The more positive the weight, the more solidly we can conclude from these sources that the assertion is true; a negative weight means we should conclude from these sources that the assertion is *not* true.

These justification structures assign a floating-point score to each assertion, representing its reliability. As such, the conjunctions and disjunctions are modeled on operators in real-valued fuzzy logic, not Boolean logic.

As in previous versions of ConceptNet, an assertion that receives a negative score is not an assertion whose negation is true. It may in fact be a nonsensical or irrelevant assertion. To represent a true negative statement, such as "Pigs cannot fly", ConceptNet 5 uses negated relations such as `/r/NotCapableOf`.

Conjunctions are necessary to assign credit appropriately to the multi-part processes that create many assertions. For example, an OMCS sentence may be typed in by a human contributor and then interpreted by a parser, and we want the ability to examine the collected data and determine whether the human is a reliable data source as well as whether the parser is. As another example, relations mined from Wikipedia using ReVerb depend on both the reliability of Wikipedia and of ReVerb.

6.2.3 Granularity

The different knowledge sources that feed ConceptNet 5 represent concepts at different levels of granularity, especially in that they can be ambiguous or disambiguated. Concepts are often ambiguous when we acquire them from natural-language text. Other concepts are explicitly disambiguated by a resource such as WordNet or Wiktionary. ConceptNet 5 contains, for example, the ambiguous node /c/en/jazz. A source such as Wiktionary might define it as a noun, yielding the more specific concept /c/en/jazz/n, and it may even distinguish the word sense from other possible senses, yielding /c/en/jazz/n/musical_art_form.

These URLs do not represent the same node, but the nodes they represent are highly related. This indicates that when we add a way to query ConceptNet 5, described in Sect. 6.3.1, we need to structure the index so that a query for /c/en/jazz also matches /c/en/jazz/n/musical_art_form.

6.2.4 Normalizing and Aligning Concepts

ConceptNet deals with natural-language data, but it should not store the assertion that "a cat is an animal" in a completely different way than "cats are animals".

Therefore, we represent each concept using *normalized* versions of the concept's text. The process for creating a normalized concept differs by language. Some examples are:

- *Running*, in English: /c/en/run
- *Rennen*, in Dutch: /c/nl/renn
- *Run (baseball)*, a disambiguated English word: `/c/en/run/n/baseball`

ConceptNet 5 uses our custom Python package called *metanl*[1] for lemmatization (reducing words to a root form) and other kinds of normalization. *metanl* provides a straightforward Python interface to our preferred stemmers and lemmatizers in many different languages.

The normalization process in English is an extension of WordNet's Morphy algorithm as provided by NLTK [3], plus removal of a very small number of stopwords, and a transformation that undoes CamelCase on knowledge sources that write their multiple-word concepts that way. In Japanese, we use the commonly-used MeCab algorithm for splitting words and reducing the words to a dictionary form [15]. In many European languages, we use the Snowball stemmer for that language [20] to remove stop words and reduce inflected words to a common stem.

Normalization inherently involves discarding information, but since ConceptNet 3, we have ensured that this information is stored with the assertion and not truly discarded. Every edge that forms every assertion is annotated with how it was expressed in natural language. That information is important in some applications such as generating natural-language questions to ask, as the AnalogySpace system [22] does with ConceptNet data; it is also very important so that if we change the normalization process one day, the original data is not lost and there is a clear way to determine which new concepts correspond to which old concepts.

6.2.5 URIs and Namespaces

An important aspect of the representation used by ConceptNet 5 is that it is free from arbitrarily-assigned IDs, such as sequential row numbers in a relational database. Every node and edge has a URI, which contains all the information necessary to identify it uniquely and no more.

Concepts (normalized terms) are the fundamental unit of representation in ConceptNet 5. Each concept is represented by a URI that identifies that it is a concept, what language it is in, its normalized text, and possibly its part of speech and disambiguation. A concept URI looks like `/c/en/run/n/basement`.

[1] http://github.com/commonsense/metanl

The predicates that relate concepts can be multilingual relations such as `/r/IsA`: this represents the "is-a" or "hypernym" relation that will be expressed in different ways, especially when the text is in different languages.

Processes that read free text, such as ReVerb, will produce relations that come from natural language and cannot be aligned in any known way with our multilingual relations. In this case, the relation is in fact another concept, with a specified language and a normalized form. In the text "A bassist performs in a jazz trio", the relation is /c/en/perform_in.

The fact that interlingual relations and language-specific concepts can be interchanged in this way is one reason we need to distinguish them with the namespaces /r/ and /c/. The namespaces are as short as possible so as to not waste memory and disk space; they appear millions of times in ConceptNet.

There is a namespace /s/ for data sources that justify an edge. These contain, for example, information extraction rules such as /s/rule/reverb, human contributors such as /s/contributor/omcs/rspeer, and curated projects such as /s/wordnet/3.0.

An assertion URI contains all the information necessary to reconstruct that assertion. For example, the assertion that "jazz is a kind of music" has the URI /a/[/r/IsA/,/c/en/jazz/,/c/en/music/]. By using the special path components /[/ and /]/, we can express arbitrary tree structures within the URI, so that the representation can even represent assertions about assertions without ambiguity. The advantage of this is that if multiple branches of ConceptNet are developed in multiple places, we can later merge them simply by taking the union of the edges. If they acquire the same fact, they will assign it the same ID.

Assertions will be represented multiple times by multiple edges. Edge IDs also take into account all the information that uniquely identifies the edge. There is no need to represent this information in a way from which its parts can be reconstructed; doing so would create very long edge IDs that would repeat the majority of the data contained in the edge. Instead, edge IDs are the hexadecimal SHA-1 hash of all the unique components, separated by spaces. These IDs can be queried to get an arbitrary subset of edges, which is very useful for evaluation.

6.2.6 Graph Statistics

A simple transformation of ConceptNet 5 allows us to consider it as a simple undirected graph. We consider there to be an edge between the two arguments of every assertion. We add an implicit edge from every disambiguated concept to its ambiguous form, and from every reified assertion to its two arguments: for example,

The resulting graph[2] has 9,611,524 distinct edges among 3,930,196 nodes.

[2]Statistics apply to the May 1, 2012 release.

The largest connected component contains 3,675,400 nodes. The second largest component, with 727 nodes, contains all the instances of /c/en/olympic_result from DBPedia, such as /c/en/belgium_at_1972_winter_olympics.

ConceptNet 5 is not overwhelmed with dangling edges; the 2-core of ConceptNet 5 (the maximal subgraph in which every node has degree ≥ 2) contains 8,286,862 edges among 2,512,028 nodes.

6.3 Storing and Accessing ConceptNet Data

As ConceptNet grows larger and is used for more purposes, it has been increasingly important to separate the data from the interface to that data. A significant problem with ConceptNet 3, for example, was that the only way to access it was through the same Django database models that created it.

ConceptNet 5 fully separates the data from the interface. The data in ConceptNet 5 is a flat list of edges, available in JSON or as tab-separated values. A flat file is in fact the most useful format for many applications:

- Many statistics about ConceptNet can be compiled by iterating over the full list of data, which neither a database nor a graph structure is optimized for.
- A subset of the information in each line of the flat file is the appropriate input for many machine learning tools.
- A flat file can be easily converted to different formats using widely-available tools.
- It is extremely easy to merge flat files. It is sometimes sufficient simply to put them in the same directory and iterate over both. If deduplication is needed, one can use highly optimized tools to sort the lines and make them unique.

However, a flat file is not particularly efficient for querying. A question such as "What are the parts of a car?" involves a very small proportion of the data, which could only be found in a flat file by iterating over the entire thing. Thus, we build indexes *on top of* ConceptNet 5.

6.3.1 Indexes

Currently, we index ConceptNet 5 with a combination of Apache Solr and MongoDB. We provide access to them through a REST API, as well as transformations of the data that a downstream user can import into a local Solr index or MongoDB database. The Solr index seems to be the most useful and scalable, and its distributed queries make it simple to distribute it between sites, so it is the primary index that we currently use. For example, we can maintain the main index while our collaborators in Taiwan maintain a separate index, including up-to-date information they have collected, and now a single API query can reach both.

Using the Solr server, we can efficiently index all edges by all lemmas (normalized words) they contain and prefixes of any URIs they involve. A search for `rel:/r/PartOf` and `end:/c/en/wheel OR end:/c/en/wheel/*` will find all edges describing the parts of a wheel, automatically ordered by the absolute value of their score. The Solr index would not make sense as a primary way to store the ConceptNet data, but it allows very efficient searches for many kinds of queries a downstream user would want to perform.

The flat file of ConceptNet 5 contains 7.5 GB of text. A Solr index performs best when it can keep all its data, plus overhead for indexing, in memory instead of swapping it to disk. This is a large memory requirement for a single computer. However, when we shard the index across two *m1.large* instances on Amazon EC2, each with 7.5 GB of RAM, the data and index fit in memory. This is sufficient to respond to queries on any of the indexed fields in 100–500 ms.

6.3.2 Downloading

ConceptNet's usefulness as a knowledge platform depends on its data being freely available under a minimally restrictive license, and not (for example) tied up in agreements to use the data only for research purposes. ConceptNet 5 can be downloaded or accessed through a Web API at its web site, http://conceptnet5.media.mit.edu, and may be redistributed or reused under a choice of two Creative Commons licenses.

The flat files containing ConceptNet 5 data are available at: http://conceptnet5.media.mit.edu/downloads/

Python code for working with this data, transforming it, and building indexes from it is maintained on GitHub in the "conceptnet5" project: https://github.com/commonsense/conceptnet5.

6.4 Evaluation

To evaluate the current content of ConceptNet, we put up a website for 48 h that showed a random sample of the edges in ConceptNet. It showed the natural language form of the text (which was machine-generated in the cases where the original data was not in natural language) and asked people to classify the statement as "Generally true", "Somewhat true", "I don't know", "Unhelpful or vague", "Generally false", and "This is garbled nonsense". People were invited to participate via e-mail and social media. They were shown 25 results at a time. We got 81 responses that evaluated a total of 1,888 statements, or 1,193 if "Don't know" results are discarded.

All participants were English speakers, so we filtered out statements whose surface text was not in English. Statements that translate another language to English were left in, but participants were not required to look them up, so in many cases they answered "Don't know".

Table 6.2 The breakdown of responses to an evaluation of random statements in ConceptNet 5

Dataset	False	Nonsense	Vague	Don't know	Sometimes	True	Total
Existing ConceptNet	34	50	15	19	117	300	535
WordNet	4	17	0	11	9	35	76
Wiktionary, English-only	2	5	3	9	6	10	35
Wiktionary, translations	4	6	2	233	8	51	304
DBPedia	10	36	9	389	41	238	723
Verbosity	10	41	7	2	32	51	143
ReVerb	2	15	15	19	3	5	59
GlobalMind translations	0	0	0	4	0	0	4
Negative edges	4	2	0	0	1	2	9

We have grouped the results by *dataset*, distinguishing edges that come from fundamentally different sources. The datasets are:

- **Existing ConceptNet**: statements previously collected by Common Sense Computing projects, which can be found in ConceptNet 4.
- **WordNet**: connections from WordNet 3.0.
- **Wiktionary, English-only**: monolingual information from the English Wiktionary, such as synonyms, antonyms, and derived words.
- **Wiktionary, translations**: translations in Wiktionary from some other language to English. As these are numerous compared to other sources, we kept only 50 % of them.
- **DBPedia**: Triples from DBPedia's `instance_types_en` dataset. As these are numerous compared to other sources, we kept only 25 % of them.
- **Verbosity**: Statements collected from players of Verbosity on gwap.com.
- **ReVerb**: Filtered statements extracted from ReVerb parses of a corpus of Wikipedia's front-paged articles.
- **GlobalMind translations**: translations of entire assertions between languages.

We also separated out **negative edges**, those which previous contributors to ConceptNet have rated as not true, confirming that most of them are rated similarly now.

The breakdown of results appears in Table 6.2. Their relative proportions, excluding the "Don't know" responses, are graphed in Fig. 6.3.

We can see that people often answered "Don't know" when faced with very specific knowledge, which is to be expected when presenting expert knowledge to arbitrary people.

All the examples of higher-level assertions that translate assertions between languages were rated as "Don't know". A more complete evaluation could be performed in the future with the help of bilingual participants who could evaluate translations.

The processes of extracting translations from Wiktionary and triples from DBPedia performed very well, while the ReVerb data – faced with the hardest task, extracting knowledge from free text – did poorly. The few negative-score edges were mostly rated as false, as expected, though 3 out of 9 of the respondents to them

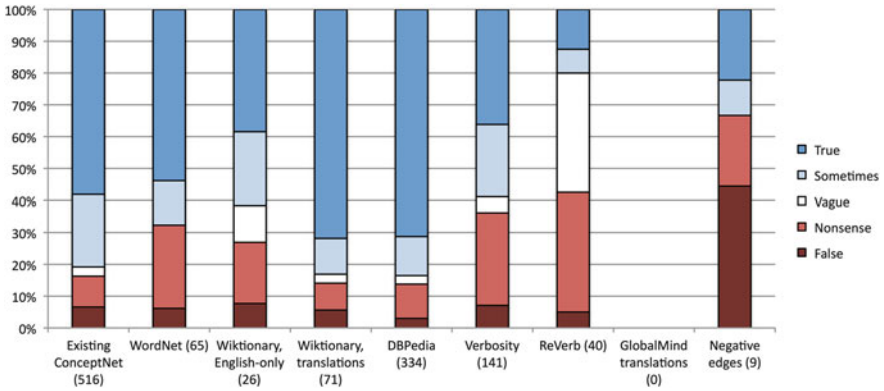

Fig. 6.3 The relative proportions of responses people gave about each dataset

disagreed. The core data in ConceptNet was evaluated nearly exactly the same as data mined from Wiktionary.

Interestingly, existing ConceptNet data was rated better than WordNet data, which was often rated as "nonsense"; perhaps the average WordNet edge is an assertion so obscure that a human evaluator will not even recognize it as making sense, or perhaps our own process of generating artificial English-language glosses of the WordNet edges is at fault.

A typical example of a WordNet edge rated "nonsense" is: *white (flesh of any of a number of slender food fishes especially of Atlantic coasts of North America)* is part of *white (any of several food fishes of North American coastal waters)*. When describing obscure senses of the word "white", this is actually a highly specific true statement, but our evaluator did not decipher it. Other statements rated "nonsense" include statements that reflect an unintuitive taxonomy, such as *illinois class battleship* is a *product* from DBPedia.

These should be distinguished from errors in which the lack of context is an inherent problem with the statement. When a process such as ReVerb extracts a statement without the context that makes it make sense, such as "*Critics* have seen *Jake*", evaluators correctly mark it as nonsense.

We present the results as they are, keeping in mind that a future evaluation should be designed to provide evaluators with more of the context they need to make accurate judgments. The presence of awkwardly-worded statements with absolutely no context had a negative effect on the evaluation of all sources, but was particularly penalizing to highly-specific statements such as those in WordNet.

6.4.1 Next Steps

The overall accuracy of approximately 79 % across all sources is sufficient for many purposes but motivates future work on verifying and cleaning up the data.

The ConceptNet data presents many starting points for machine learning, which could help to both refine the ConceptNet data and to create new resources from it. The ConceptNet 5 Web API[3] currently supports using dimensionality reduction, as in [22], to list similar concepts to a query. Useful future tasks include automatically learning from and refining the assertion scores to learn which sources and combinations of sources provide the most reliable information, aligning the most similar word senses within a language and across different languages, and recognizing paraphrases and nearly-equivalent statements that support one another.

References

1. Anacleto J, Lieberman H, Tsutsumi M, Neris V, Carvalho A, Espinosa J, Zem-Mascarenhas S (2006) Can common sense uncover cultural differences in computer applications? In: Proceedings of IFIP world computer conference, Santiago, Chile
2. Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z (2007) DBpedia: a nucleus for a web of open data. In: Aberer K, Choi KS, Noy N, Allemang D, Lee KI, Nixon L, Golbeck J, Mika P, Maynard D, Mizoguchi R, Schreiber G, Cudré-Mauroux P (eds) The semantic web. Lecture notes in computer science, vol 4825. Springer, Berlin/Heidelberg, pp 722–735
3. Bird S, Klein E, Loper E (2009) Natural language processing with Python. O'Reilly Media, Beijing
4. Blanco E, Cankaya HC, Moldovan D (2011) Commonsense knowledge extraction using concepts properties. In: Proceedings of the 24th Florida artificial intelligence research society conference (FLAIRS-24), Palm Beach, FL, USA, pp 222–227
5. Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J (2008) Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on management of data, SIGMOD '08. ACM, New York, pp 1247–1250. doi:http://doi.acm.org/10.1145/1376616.1376746
6. Cambria E, Hussain A, Havasi C, Eckl C (2010) SenticSpace: visualizing opinions and sentiments in a multi-dimensional vector space. In: Knowledge-based and intelligent information and engineering systems. Springer, Heidelberg, pp 385–393
7. de Melo G, Weikum G (2010) Menta: inducing multilingual taxonomies from Wikipedia. In: Proceedings of the 19th ACM international conference on information and knowledge management, CIKM '10. ACM, New York, pp 1099–1108. doi:http://doi.acm.org/10.1145/1871437.1871577
8. Eckhardt N (2008) A kid's open mind common sense. PhD thesis, Tilburg University, http://ilk.uvt.nl/downloads/pub/papers/GV_thesis_NienkeFINAL.pdf
9. Etzioni O, Banko M, Soderland S, Weld DS (2008) Open information extraction from the web. Commun ACM 51:68–74. doi:http://doi.acm.org/10.1145/1409360.1409378
10. Fellbaum C (1998) WordNet: an electronic lexical database. MIT, Cambridge, MA
11. Havasi C, Speer R, Alonso J (2007) ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge. In: recent advances in natural language processing, Borovets, Bulgaria, pp 27–29. http://web.mit.edu/~rspeer/www/research/cnet3.pdf
12. Havasi C, Borovoy R, Kizelshteyn B, Ypodimatopoulos P, Ferguson J, Holtzman H, Lippman A, Schultz D, Blackshaw M, Elliott G, Ng C (2011) The glass infrastructure: using common sense to create a dynamic, place-based social information system. In: Proceedings of 2011 conference on innovative applications of artificial intelligence. AAAI, San Francisco

[3] https://github.com/commonsense/conceptnet5/wiki/API

13. Kochenderfer MJ (2004) Common sense data acquisition for indoor mobile robots. In: Proceedings of the nineteenth national conference on artificial intelligence (AAAI-04), San Jose, California, at the San Jose Convention Center, July 25–29, 2004, pp 605–610
14. Korner S, Brumm T (2009) Resi – a natural language specification improver. In: IEEE international conference on semantic computing, 2009, ICSC '09, pp 1–8. doi:10.1109/ICSC.2009.47
15. Kudo T, Yamamoto K, Matsumoto Y (2004) Applying conditional random fields to Japanese morphological analysis. In: Lin D, Wu D (eds) Proceedings of EMNLP 2004. Association for Computational Linguistics, Barcelona, pp 230–237
16. Kuo YL, Lee JC, Chiang KY, Wang R, Shen E, Chan CW, Hsu JYJ (2009) Community-based game design: experiments on social games for commonsense data collection. In: Proceedings of the ACM SIGKDD workshop on human computation, HCOMP '09. ACM, New York, pp 15–22, doi:http://doi.acm.org/10.1145/1600150.1600154
17. Nastase V, Strube M, Boerschinger B, Zirn C, Elghafari A (2010) Wikinet: a very large scale multi-lingual concept network. In: LREC, Valletta, Malta, May 17–23, 2010
18. Navigli R, Ponzetto SP (2010) Babelnet: building a very large multilingual semantic network. In: Proceedings of the 48th annual meeting of the association for computational linguistics, ACL '10. Association for Computational Linguistics, Stroudsburg, pp 216–225. http://dl.acm.org/citation.cfm?id=1858681.1858704
19. Niles I, Pease A (2001) Towards a standard upper ontology. In: Proceedings of the international conference on formal ontology in information systems – volume 2001, FOIS '01. ACM, New York, pp 2–9. doi:http://doi.acm.org/10.1145/505168.505170
20. Porter MF (2001) Snowball: a language for stemming algorithms. Published online at http://snowball.tartarus.org/texts/introduction.html. Accessed 24 Oct 2011
21. Singh P, Lin T, Mueller ET, Lim G, Perkins T, Zhu WL (2002) Open Mind Common Sense: knowledge acquisition from the general public. In: On the move to meaningful internet systems, 2002 – DOA/CoopIS/ODBASE 2002 Confederated international conferences DOA, CoopIS and ODBASE 2002. Springer, London, pp 1223–1237. http://www.media.mit.edu/~push/ODBASE2002.pdf
22. Speer R, Havasi C, Lieberman H (2008) AnalogySpace: reducing the dimensionality of common sense knowledge. In: Proceedings of AAAI, Chicago, Illinois, July 13-17, 2008
23. Speer R, Havasi C, Treadway N, Lieberman H (2010) Finding your way in a multi-dimensional semantic space with Luminoso. In: Proceedings of the 15th international conference on intelligent user interfaces, Hong Kong, China, February 7–10, 2010
24. Ullberg J, Coradeschi S, Pecora F (2010) On-line ADL recognition with prior knowledge. In: Proceeding of the 2010 conference on STAIRS 2010: proceedings of the fifth Starting AI Researchers' symposium. IOS, Amsterdam, pp 354–366. http://dl.acm.org/citation.cfm?id=1940526.1940556
25. van Assem M, Isaac A, von Ossenbruggen J (2010) Wordnet 3.0 in RDF. Published online at http://semanticweb.cs.vu.nl/lod/wn30/. Accessed 24 Oct 2011
26. von Ahn L, Kedia M, Blum M (2006) Verbosity: a game for collecting common-sense facts. In: Proceedings of the SIGCHI conference on human factors in computing systems, CHI '06. ACM, New York, pp 75–78. doi:http://doi.acm.org/10.1145/1124772.1124784
27. Zesch T, Müller C, Gurevych I (2008) Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In: Proceedings of the 6th conference on language resources and evaluation (LREC). http://elara.tk.informatik.tu-darmstadt.de/publications/2008/lrec08_camera_ready.pdf