Representing General Relational Knowledge in ConceptNet 5

Robert Speer and Catherine Havasi

MIT Media Lab
20 Ames St., Cambridge MA 02139
rspeer@mit.edu, havasi@mit.edu

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. This paper presents the latest iteration, ConceptNet 5, including its fundamental design decisions, ways to use it, and evaluations of its coverage and accuracy.

Keywords: ConceptNet, common sense, semantic networks

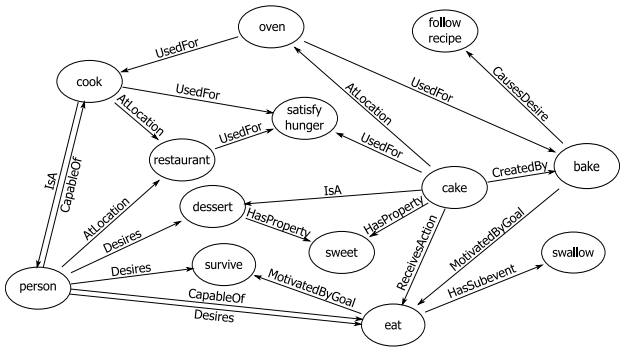

Figure 1: A high-level view of the knowledge ConceptNet has about a cluster of related concepts.

1. Introduction

ConceptNet is a knowledge representation project, 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, describing not just how words are related by their lexical definitions, but also how they are related through common knowledge. For example, its 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 assertions that connect them is visualized in Figure 1.

ConceptNet originated as a representation for the knowledge collected by the Open Mind Common Sense project (Singh et al., 2002), 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.

ConceptNet gives a foundation of real-world knowledge to a variety of AI projects and applications. Previous versions of ConceptNet (Havasi et al., 2007) have been used, for example, to build a system for analyzing the emotional content of text (Cambria et al., 2010), to create a dialog system for improving software specifications (Korner and Brumm, 2009), to recognize activities of daily living (Ullberg et al., 2010), to visualize topics and trends in a corpus of unstructured text (Speer et al., 2010), and to create public information displays by reading text about people and projects from a knowledge base (Havasi et al., 2011).

ConceptNet provides a combination of features not available in other knowledge representation projects:

- Its concepts are connected to 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.
- 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.

2. Motivation for ConceptNet 5

The new goals of ConceptNet 5 include to include knowledge from other crowd-sourced knowledge with their own communities and editing processes, particularly data mined from Wiktionary and Wikipedia; to add links to other resources such as DBPedia (Auer et al., 2007), Freebase (Bollacker et al., 2008), and WordNet (Fellbaum, 1998); to support machine-reading tools such as ReVerb (Etzioni et al., 2008), which extracts relational knowledge from Web pages; and to find translations between concepts represented in different natural languages.

ConceptNet 5 aims 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 its collected knowledge when it 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 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.

2.1. Previous Development of ConceptNet

ConceptNet has been developed as part of the Open Mind Common Sense project, a Media Lab project to collect the things that computers should know in order to understand what people are talking about, which then grew into an international, multi-homed project called the Common Sense Computing Initiative.

The first publicly released version of ConceptNet was ConceptNet 2 (Liu and Singh, 2004). ConceptNet 2 was distributed as a packed Python data structure, along with code to read it and operations that could be performed with it such as spreading activation from a set of words.

ConceptNet 2 was only in English. The project became multilingual shortly afterward, both with the sister project OMCS no Brasil (Anacleto et al., 2006), collecting knowledge in Brazilian Portuguese, and GlobalMind (Chung, 2006), collecting knowledge in English, Chinese, Japanese, and Korean.

ConceptNet 3 (Havasi et al., 2007) made ConceptNet into a SQL database, so it could be easily updated by processes including user interaction from a Web site. There were separate editions of ConceptNet 3 for English and Brazilian Portuguese.

ConceptNet 4 was quite similar to ConceptNet 3, but it revised and normalized the database structure so that it could contain all languages of ConceptNet simultaneously, so that it finally could represent all the knowledge from the English OMCS, OMCS no Brasil, a new OMCS in Dutch (Eckhardt, 2008), and GlobalMind in one place. It also incorporated contributions from other projects, including online games collecting knowledge in English, Chinese, and Japanese. To aid the use of ConceptNet within other projects, we also added a Web API for accessing and querying the data in ConceptNet 4.

A strong motivation for why a new version is necessary is that the data from other projects was difficult to fully incorporate into ConceptNet 4. It had to be aligned and deduplicated in a sprawling SQL database, a time-consuming and code-intensive process that was performed infrequently, and meanwhile the projects at other locations had to maintain their own out-of-sync versions of the database. ConceptNet 5 contains many representational improvements, but the primary focus is to make the collection, storage, and querying of knowledge truly distributable.

3. Knowledge in ConceptNet 5

ConceptNet expresses *concepts*, which are words and phrases that can be extracted from natural language text, and *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 (Anacleto et al., 2006) and Dutch (Eckhardt, 2008).
- The multilingual data, including translations between assertions, collected by GlobalMind.
- "Games with a purpose" that collect common knowledge, including Verbosity (von Ahn et al., 2006) in English, *nadya.jp* in Japanese, and the "pet game" (Kuo et al., 2009) 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.
- WordNet 3.0 (Fellbaum, 1998), including cross-references to its RDF definition at `http://semanticweb.cs.vu.nl/lod/wn30/` (van Assem et al., 2010).
- The semantic connections between Wikipedia articles represented in DBPedia (Auer et al., 2007), 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 "instance_types_en" collection of DBPedia.
- Relational statements mined from Wikipedia's free text using ReVerb (Etzioni et al., 2008), 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.

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 (Kochenderfer, 2004), connections to knowledge in Freebase (Bollacker et al., 2008), ontological connections to SUMO and MILO (Niles and Pease, 2001), and new processes that scan well-structured Wiktionaries in other target languages, such as Japanese and German.

3.1. Representation

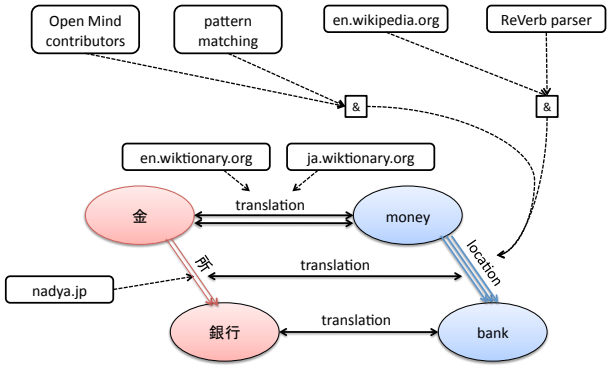

Figure 2: An example of two assertions in ConceptNet 5, and the edges they involve. Rounded rectangles and dotted edges represent knowledge sources; solid edges are grouped together into assertions.

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). 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 "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 Figure 2.

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 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.

The sources that justify each assertion form a disjunction of conjunctions. Each edge – that is, each instance – indicates a conjunction of sources that produced that edge, while the bundle of edges making up an assertion represents the disjunction of all those conjunctions. Each conjunction comes with a positive or negative *score*, a weight that it as-

signs to that edge. 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.

As in previous versions of ConceptNet, an assertion that receives a negative weight 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.

3.2. Relations

In addition to free-text relations, the standard interlingual relations we identify in ConceptNet appear in Table 1.

3.3. Granularity

The different knowledge sources that feed ConceptNet 5 represent concepts at different levels of granularity, especially in that concepts 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 `/concept/en/jazz`. A source such as Wiktionary might define it as a noun, yielding the more specific concept `/concept/en/jazz/n`, and it may even distinguish the word sense from other possible senses, yielding `/concept/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 Section 4.1., we need to structure the index so that a query for `/concept/en/jazz` also matches `/conceptnet/en/jazz/n/musical_art_form`.

3.4. 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`
- *correr*, in Spanish: `/c/es/corr`
- *rennen*, in Dutch: `/c/nl/renn`
- *run (baseball)*, a disambiguated English word: `/c/en/run/n/baseball`

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 (Speer et al., 2008) 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.

The normalization process in English is an extension of WordNet's Morphy, 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 (Kudo et al., 2004), and in many European languages we use the Snowball stemmer for that language (Porter, 2001) to remove stop words and reduce inflected words to a common stem.

3.5. 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`.

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 extrac-

Relation	Sentence pattern	Relation	Sentence pattern
IsA	*NP* is a kind of *NP*.	LocatedNear	You are likely to find *NP* near *NP*.
UsedFor	*NP* is used for *VP*.	DefinedAs	*NP* is defined as *NP*.
HasA	*NP* has *NP*.	SymbolOf	*NP* represents *NP*.
CapableOf	*NP* can *VP*.	ReceivesAction	*NP* can be *VP*.
Desires	*NP* wants to *VP*.	HasPrerequisite	*NP*|*VP* requires *NP*|*VP*.
CreatedBy	You make *NP* by *VP*.	MotivatedByGoal	You would *VP* because you want *VP*.
PartOf	*NP* is part of *NP*.	CausesDesire	*NP* would make you want to *VP*.
Causes	The effect of *VP* is *NP*|*VP*.	MadeOf	*NP* is made of *NP*.
HasFirstSubevent	The first thing you do when you *VP* is *NP*|*VP*.	HasSubevent	One of the things you do when you *VP* is *NP*|*VP*.
AtLocation	Somewhere *NP* can be is *NP*.	HasLastSubevent	The last thing you do when you *VP* is *NP*|*VP*.
HasProperty	*NP* is *AP*.		

Table 1: The interlingual relations in ConceptNet, with example sentence frames in English.

tion 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 even includes assertions about assertions. 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.

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, which are unnecessary because edges are the lowest level of data the parts of every edge are right there in the edge's data structure. Instead, edge IDs are the hexadecimal SHA-1 hash of all the unique components, separated by spaces: its assertion URI, its context, and its conjoined sources in Unicode sorted order. The 160-bit SHA-1 hash provides more than enough room to be unique over even a large number of edges, is shorter than the data contained in the edge itself, and can be queried to get an arbitrary subset of edges, which is very useful for evaluation.

4. 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.
- A CSV flat file can be used as a spreadsheet.
- 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.

4.1. 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.

4.2. 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`.

5. Evaluation

To evaluate the current content of ConceptNet, we put up a website for 48 hours 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 1888 statements, or 1193 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".

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 Initiative 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 2. Their relative proportions are graphed in Figure 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. Interestingly, existing ConceptNet data was rated better than WordNet data; perhaps WordNet edges inherently form assertions that sound too unnatural, or perhaps our English-language glosses of them are at fault. 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 disagreed.

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.

6. References

Junia Anacleto, Henry Lieberman, Marie Tsutsumi, Vania Neris, Aparecido Carvalho, Jose Espinosa, and Silvia Zem-Mascarenhas. 2006. Can common sense uncover cultural differences in computer applications? In *Proceedings of IFIP World Computer Conference*, Santiago, Chile.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer Berlin / Heidelberg.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 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, pages 1247–1250, New York, NY, USA. ACM.

Erik Cambria, Amir Hussain, Catherine Havasi, and Chris Eckl. 2010. SenticSpace: Visualizing opinions and sentiments in a multi-dimensional vector space. In *Knowledge-Based and Intelligent Information and Engineering Systems*, pages 385–393, Heidelberg.

Dataset	False / Nonsense	Vague	Don't know	Sometimes	True	Total
Existing ConceptNet	84	15	19	117	300	535
WordNet	21	0	11	9	35	76
Wiktionary, English-only	7	3	9	6	10	35
Wiktionary, translations	10	2	233	8	51	304
DBPedia	46	9	389	41	238	723
Verbosity	51	7	2	32	51	143
ReVerb	17	15	19	3	5	59
GlobalMind translations	0	0	4	0	0	4
Negative edges	6	0	0	1	2	9

Table 2: The breakdown of responses to an evaluation of random statements in ConceptNet 5.

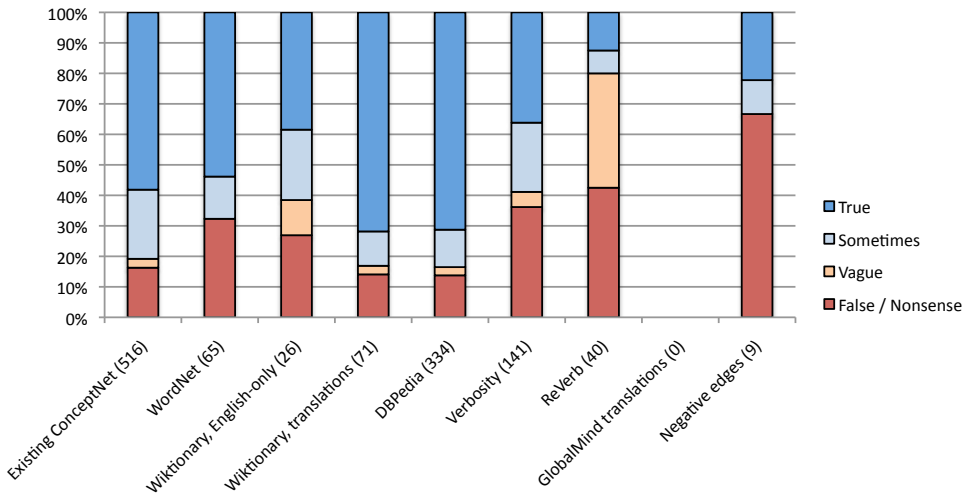

Figure 3: The relative proportions of responses people gave about each dataset.

Hyemin Chung. 2006. *GlobalMind — bridging the gap between different cultures and languages with commonsense computing*. Ph.D. thesis, MIT Media Lab.

Nienke Eckhardt. 2008. *A Kid's Open Mind Common Sense*. Ph.D. thesis, Tilburg University.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Commun. ACM*, 51:68–74, December.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Catherine Havasi, Robert Speer, and Jason Alonso. 2007. ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent Advances in Natural Language Processing*, pages 27–29, Borovets, Bulgaria, September.

Catherine Havasi, Richard Borovoy, Boris Kizelshteyn, Polychronis Ypodimatopoulos, Jon Ferguson, Henry Holtzman, Andrew Lippman, Dan Schultz, Matthew Blackshaw, Greg Elliott, and Chaki Ng. 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*, San Francisco, CA, July. AAAI.

Mykel J. Kochenderfer. 2004. Common sense data acquisition for indoor mobile robots. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, pages 605–610.

S.J. Korner and T. Brumm. 2009. Resi - a natural language specification improver. In *Semantic Computing, 2009. ICSC '09. IEEE International Conference on*, pages 1 – 8, sept.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 230–237, Barcelona, Spain, July. Association for Computational Linguistics.

Yen-Ling Kuo, Jong-Chuan Lee, Kai-Yang Chiang, Rex Wang, Edward Shen, Cheng-Wei Chan, and Jane Yung-Jen Hsu. 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, pages 15–22, New York, NY, USA. ACM.

Hugo Liu and Push Singh. 2004. ConceptNet: A practical commonsense reasoning toolkit. *BT Technology Journal*, 22(4):211–226, October.

Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001*, FOIS '01, pages 2–9, New York, NY, USA. ACM.

Martin F. Porter. 2001. Snowball: A language for stemming algorithms. Published online at http://snowball.tartarus.org/texts/introduction.html, October. Accessed 2011-10-24.

Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan L. Zhu. 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*, pages 1223–1237, London, UK. Springer-Verlag.

Robert Speer, Catherine Havasi, and Henry Lieberman. 2008. AnalogySpace: Reducing the dimensionality of common sense knowledge. *Proceedings of AAAI 2008*, July.

Robert Speer, Catherine Havasi, Nichole Treadway, and Henry Lieberman. 2010. Finding your way in a multi-dimensional semantic space with Luminoso. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*.

Jonas Ullberg, Silvia Coradeschi, and Federico Pecora. 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*, pages 354–366, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Mark van Assem, Antoine Isaac, and Jacco von Ossenbruggen. 2010. Wordnet 3.0 in RDF. Published online at http://semanticweb.cs.vu.nl/lod/wn30/, September. Accessed 2011-10-24.

Luis von Ahn, Mihir Kedia, and Manuel Blum. 2006. Verbosity: a game for collecting common-sense facts. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, CHI '06, pages 75–78, New York, NY, USA. ACM.