Verbal IQ of a Four-Year Old Achieved by an AI System

Stellan Ohlsson and **Robert H. Sloan**
University of Illinois at Chicago
stellan | sloan @uic.edu

György Turán
University of Illinois at Chicago
University of Szeged
gyt@uic.edu

Aaron Urasky
University of Illinois at Chicago
aaron.urasky@gmail.com

Abstract

One view of common-sense reasoning ability is that it is the ability to perform those tasks with verbal inputs and outputs that have traditionally been difficult for computer systems, but are easy for fairly young children. We administered the verbal part of the Wechsler Preschool and Primary Scale of Intelligence (WPPSI-III, Third Edition) to the ConceptNet 4 system.

The IQ test's questions (e.g., "Why do we shake hands?" or "What do apples and bananas have in common") were translated into ConceptNet 4 inputs using a combination of the simple natural language processing tools that come with ConceptNet together with short Python programs that we wrote. The question-answering primarily used the part of the ConceptNet system that represents the knowledge as a matrix based on spectral methods (AnalogySpace).

We found that the system has a Verbal IQ that is average for a four-year-old child, but below average for 5, 6, and 7 year-olds. Large variations from subtest to subtest indicate potential areas of improvement. In particular, results were strongest for the Vocabulary and Similarities subtests, intermediate for the Information subtest, and lowest for the Comprehension and Word Reasoning subtests. Comprehension is the subtest most strongly associated with common sense.

Children's verbal IQ tests offer a new, objective, third-party metric for the evaluation and comparison of common-sense AI systems.

Introduction

The question how computer programs might exhibit common sense was implicit in Turing's work (Turing 1950), and explicit in McCarthy's seminal 1959 AI paper, "Programs with Common Sense" (McCarthy 1959). At the beginning of his 1990 book on common-sense knowledge, Davis defines common sense as "common knowledge about the world that is possessed by every schoolchild and the methods for making obvious inferences from this knowledge" (Davis 1990), and illustrates it with an example "easily understood by five-year-old children." In the closing chapter of his 2010 account of the history of AI, Nilsson (Nilsson 2009) quotes a list of challenges posed by Rodney Brooks (Brooks 2008): the object-recognition capabilities of a two-year old child, the language capabilities of a four-year old child, the manual dexterity of a six-year-old child and the social understanding of an eight-year-old child.

Capturing language capabilities, social understanding, and common sense in a computer system has turned out to be even more difficult than capturing technical expertise. Reasoning of this sort appears to draw upon a factual and conceptual knowledge base of vast proportions. So far, researchers have not found efficient, effective, and unified implementations for many types of inferences people easily engage in, such as counterfactual reasoning or reasoning about others' mental states. The logical formalization of these types of reasoning has been studied intensively in the past two decades, e.g., (Fagin et al. 1995; Gärdenfors 1988). One approach to this problem is to invest the large resources required to create a knowledge base that matches the knowledge base of a human being, in the hope that once the computer has all the relevant knowledge, it, too, will exhibit common sense.

In recent decades researchers have come to realize that such a large knowledge base of facts is probably a prerequisite for successful common-sense reasoning. Attempts to build such a knowledge base include ConceptNet/AnalogySpace (Speer, Havasi, and Lieberman 2008; Havasi, Speer, and Alonso 2007; Havasi et al. 2009), Cyc (Lenat 1995), and Scone (Fahlman 2006). Two of these systems, ConceptNet and Cyc, have only become publicly available relatively recently.

Lacking a generally accepted performance standard, it is impossible to evaluate claims and document progress of

these systems. We set out to measure at least one such system by using a test of intelligence developed by psychometricians: the IQ test. We used the Wechsler Preschool and Primary Scale of Intelligence, Third Edition (WPPSI-III) test, a multi-dimensional IQ test designed to assess the intelligence of children of ages 2.5–7.25 years (Wechsler 2002), although many parts of it can be used only with children at least 4 years old. The WPPSI-III is one of two commonly used IQ tests for young children; the other is the Stanford-Binet. There are also various IQ tests for older children and adults.

Although there is a long-standing debate regarding the usefulness of intelligence tests for managing societal affairs (Perkins 1995), the present study uses these tests as a standardized performance measure of computer systems, a usage that does not directly engage that debate.

The WPPSI-III questions are proprietary, so as in all scientific work reporting on IQ testing, we will not report on the specific questions in that test. In order to explore whether using such a test was even possible, we first made up our own test items in the general spirit of, but distinct from, the WPPSI-III. Specific examples of questions we give come from those, although the reported scores are from the actual WPPSI-III, used under license. We reported on our planned methodology and some preliminary results based on our sample questions in (Ohlsson et al. 2012).

We initially attempted to explore Cyc in addition to ConceptNet, but we were unable to use Cyc to answer more than a few of our sample questions. This may reflect the difficulty of learning to use Cyc, rather than any limitations on Cyc's abilities. We were successful in using ConceptNet 4 and the associated AnalogySpace (specifically the version provided in Python by the divisi package) to answer a number of our sample questions.

We report here on ConceptNet's performance on the WPPSI-III Verbal IQ (VIQ). As required by the test, grading was done by a psychologist (author Ohlsson). The VIQ results are necessarily a function of both ConceptNet and our algorithms that use ConceptNet to answer the questions.

While the specific questions making up the WPPSI-III are proprietary, the set of subtests and the general nature of each subtest is widely available public information.. We selected the five subscales of the test for which the items could be input into a computer system with a relatively direct translation (ruling out, e.g., subtests where the child has to work with blocks or draw). These turn out to be precisely the five subtests from which the verbal IQ is defined.

The WPPSI-III Exam

A WPPSI-III full-scale IQ is determined by a Performance IQ (drawing, puzzle, and memory tasks) and a Verbal IQ (VIQ); the IQs have mean 100 and standard deviation 15. VIQ is determined by three of the five verbal subtests: Information, Vocabulary, Word Reasoning, Comprehension, and Similarities, of which the first three are "core" and the last two "supplemental." The examiner may choose any three subtests with the constraint that at least two must be core. Subtests have mean 10, standard deviation 3.

In a Vocabulary item, the testee is asked to articulate the meaning of a common word, using the question frame, "What is ___?", like "What is a house?" Performance on a vocabulary item requires retrieval of a definition of the given concept, in combination with a lack of retrieval of irrelevant concept definitions.

In an Information item, the testee is asked to state the properties, kind, function, cause, origin, consequence, location, or other aspect of some everyday object, event, or process. For example, the testee might be asked, "Where can you find a penguin?"

In a Similarities item, two words have to be related using the sentence frame, "Finish what I say. X and Y are both ___", like "Finish what I say. Pen and pencil are both ___". Performance on a Similarities item requires the retrieval of the two concept definitions (meanings), plus the ability to find a meaningful overlap between them.

In a Word Reasoning item, the task is to identify a concept based on one to three clues. The testee might be told, "You can see through it," as a first clue; if the correct answer is not forthcoming, the testee might be told additionally that, "It is square and you can open it." The processing required by a Word Reasoning items goes beyond retrieval because the testee has to integrate the clues and choose among alternative hypotheses.

Finally, in a Comprehension item, the task is to produce an explanation in response to a why-question. The testee might be asked, "Why do we keep ice cream in the freezer?" Performance on a comprehension item requires the construction of an explanation, and so goes beyond retrieval. In descriptions of the WPPSI-III, Comprehension is often described as being a test of "common sense" as opposed to "reasoning" or some other cognitive ability.

ConceptNet/AnalogySpace

ConceptNet is an open-source project run by the MIT Common Sense Computing Initiative. It has several components. The Open Mind Common Sense initiative acquired a large common-sense knowledge base from web users (Singh 2002). This is ConceptNet itself, consisting of

triples of the form (<concept1>, relation, <concept2>), where relation is drawn from a fixed set of about two dozen relations such as IsA, HasA, UsedFor, CapableOf, HasProperty, Causes, and AtLocation. The full list is available in the documentation at http://csc.media.mit.edu/docs/conceptnet/. Henceforth, all mentions of ConceptNet refer to ConceptNet 4, specifically the version released in March 2012.

More precisely, each entry in ConceptNet consists of two concepts and one of the relations, together with either "left" or "right" to show the direction of the relation (e.g., to indicate that "a fawn IsA deer" as opposed to "a deer IsA fawn") and a numerical strength, and a polarity flag, which is set in a small minority of cases (3.4 percent) to indicate negation (e.g., polarity could be used to express the assertion that "Penguins are *not* capable of flying."). (There is also a frequency, which we did not use in this work.)

AnalogySpace is a concise version of the knowledge base (Speer, Havasi, and Lieberman 2008; Havasi et al. 2009). Leaving out assertions that have little support shrinks the number of so-called "concepts" from roughly 275,000 to roughly 22,000 for the English-language version. Additional shrinkage comes from treating the ConceptNet knowledge base as a large but sparse matrix and applying spectral techniques, specifically a truncated singular value decomposition (SVD) to obtain a smaller, denser matrix. This reduced-dimension matrix, which is called AnalogySpace, is claimed to give better, more meaningful descriptions of the knowledge.

ConceptNet is a hybrid between logical (i.e., symbolic) and statistical systems. Its triples form a classic semantic net, but the SVD is related to Principal Component Analysis.

Our Methods

Our objective was to explore the capabilities available in ConceptNet for a non-expert user without a significant knowledge engineering effort; extending this endeavor to more sophisticated algorithms is an important topic for future work. Thus, we wrote fairly short programs in Python to feed each of the five types of items into ConceptNet. We used the fairly rudimentary natural language processing tools that come with ConceptNet, and added some additional minimal natural language processing of our own. We used the sample questions we made up to develop our method, and to choose the amount of truncation of the SVD. (We settled on truncated to the first $k = 500$ most significant eigenvalues, but our results were fairly similar in quality for any value of k in the range of roughly 200 to 600.)

Our Methodology for Querying ConceptNet

We describe our methodology for querying vocabulary items in some detail, and then give somewhat shorter descriptions for the rest of the subtests, highlighting new issues these subtests raise.

Vocabulary Questions

Vocabulary questions: Our program's input is the single word being queried, for example, "house." We used the following procedure:

1. Use ConceptNet's natural language tools to map the input word to a concept in the system.
2. Query AnalogySpace for its top-scoring entry for that concept that uses one of the relations: 'IsA', 'HasA', 'HasProperty', 'UsedFor', 'CapableOf', 'DefinedAs', 'MotivatedByGoal', or 'Causes', restricting to cases where the query item is on the proper side of the relation (e.g., we want to consider only that "Sun IsA star'" and not that "Alpha Centauri IsA sun."). Additionally, we make sure that the item we just chose is not the input word itself. That entry will be a "feature," which is to say a relation together with a direction and a concept. The direction tells whether the feature is, for example "star IsA" or "IsA star".
3. For that top AnalogySpace feature, find the top-scored assertion in ConceptNet using the same pair of concepts (and typically but not necessarily the same relation).
4. For that ConceptNet assertion, find the top-scored "Raw Assertion." Raw assertions are very lightly processed user inputs from the original Open Mind learning phase.
5. Finally, apply a ConceptNet natural language function to translate that raw assertion back into English.

In the case of house, the top entry from AnalogySpace we get in Step 2 relates house to windows, not what we want; in the case of airplane, the top entry from AnalogySpace we get in Step 2 relates airplane to travel, a good answer.

At the end of the process, the top few answers for house are: "houses usually have windows;" "a house has a yard;" and "the house was probably expensive." For airplane we get: "airplanes are used to travel;" "An airplane is a form of transportation;" and "You can use an airplane to fly."

(Some additional Python code in Step 2 ensures that when we find further answers in addition to the top-scoring answer, that those answers are not too close, but rather at an appropriate "spreading activation" distance from answers we have already found.)

Information and Comprehension Questions

For both Information and Comprehension we use exactly the same procedure, and our input is the entire natural language question, such as "What color is the sky?"

First we take note of the beginning of the question, which will guide our response. If the question begins specifically "What color is/are" or "How many," then we pass it off to special subroutines we wrote for those questions. These treat the rest of the question as a bag of words, using ConceptNet's natural language tools, and return the highest scoring AnalogySpace item for that bag of words that is a color or number respectively.

Otherwise, if the beginning of the question is in our list of common beginnings, we remember it (to select relevant relations later), and remove it. The common beginnings include: "why", "where," 'what,' "why are some people," "tell me the names," "name" and several variants.

We feed the remaining words of question into the natural language tools of ConceptNet, which will remove common stop words, and return a list of ConceptNet concepts. We remove from this list of concepts any one-word concepts that are part of a two-word concept in our list that the version of AnalogySpace we are working with has entries for. For example, if our list was ['shake', 'hand", 'shake hand'] and 'shake hand' was a concept that AnalogySpace has, then we would remove both 'shake' and 'hand'.

We then create an AnalogySpace category from those concepts, which can be thought of as a column vector of concepts.

Next we take the product of the entire AnalogySpace matrix and that column vector to get a vector of proposed answers.

We return the top-scoring answer, with the restrictions that:

- The beginning of the sentence, if it was on our list of special question beginnings, will restrict the relations we consider. For example, for where questions, we considered only the two relations AtLocation and LocatedNear.
- We also look for a limited number of close matches to relation names elsewhere in the sentence; for example, for the question, "What are pancakes made out of?", we restrict to the MadeOf relation because the question contains the phrase "made out of".

Otherwise, we are using ConceptNet for question answering precisely as proposed in its documentation and tutorials.

For Information and Comprehension items, there is no obvious way to translate they system's answer back to a good English sentence. Rather our answer is a feature. For example,"MadeOf flour" is the (correct) answer we obtain to "What are pancakes made out of?"

Word Reasoning Questions

We use essentially the same procedure as for Information and Comprehension. Here we have no special treatment of the beginning of the sentence (which typically would not be relevant for these questions anyway). We do, however, after translating to concepts, remove some very common concepts that proved to be unhelpful The concepts we removed are: 'person', 'get', 'need', 'make', 'out', 'up', 'often', 'look', 'not', 'keep', 'see', and 'come'. (The removal of these words helped some on our made-up test questions, but may not have made any difference at all on the actual WPPSI-III questions.)

For second and third clues, we simply add them to the input to the ConceptNet natural language tools.

Similarity Questions

For similarities, our inputs are two words, such as *snake* and *alligator*.

For each word, we found the concept for the word and its two closest neighbors using the "spreading activation" function of AnalogySpace, and for each of those six concepts, we find the 100 highest rated features and their scores.

Using AnalogySpace, we create a set of scored predicted features for each word. Each set could have up to 300 entries, though typically both sets have many fewer, since we expect many common entries among a concept and its two closest neighbors.

We then find the features in the intersection of the two sets, and return as our answer the highest scored feature, where we are determining score by adding the score from each set.

Scoring

In all cases we got answers from the system with a score showing, intuitively, the system's degree of belief in that answer. We scored the WPPSI-III subtests once using the top-scored answer to each test item, and again using the best answer from among the five top-scoring items. We call the former *strict* and the latter *relaxed*. The relaxed score gives some idea of how heavily the system's assigned degree-of-belief weights affected the results.

Results

Raw scores, strict and relaxed, are given for each of the five subtests in Table 1. VIQ for ConceptNet, using the standard choice of three subtests (Information, Verbal, and Word Reasoning) as a function of age in years are shown in Figure 1. Relaxed scoring leads to only slightly higher scores; the only large difference is for the Similarity subtest.

	Subtests included in VIQ			Scoring Regimen	
Subtest:	Standard	Best 3	Worst 3	Strict	Relaxed
Information	x	x	x	20	21
Word Reasoning	x		x	3	3
Vocabulary	x	x		20	21
Similarities		x		24	37
Comp.			x	2	2

Table 1. *Raw WPPSI-III verbal subscores obtained with the ConceptNet system, using two different scoring regiments (see text). Also shows which subscores are used in which computations of VIQ reported in the text and in Figure 2.*

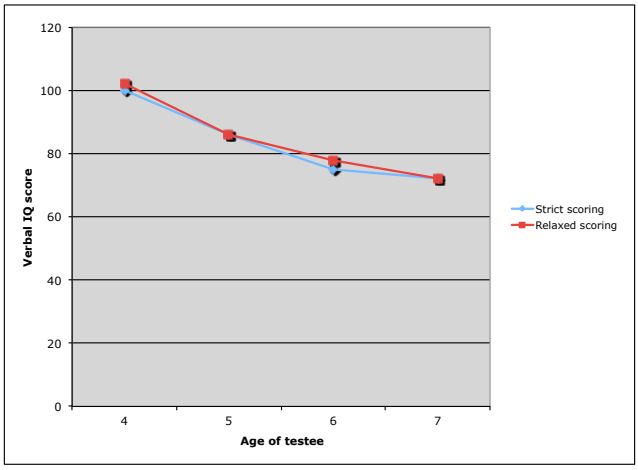

Figure 1. *WPPSI-III VIQ of ConcepteNet as a function of assumed age in years, computed using the three standard subtests, suing both strict and relaxed scoring.*

If we assume the testee is 4 years old, the VIQ score is average (VIQ = 100, based on subscores Information 10, Vocabulary 13, Word Reasoning 7) but at an assumed age of 5 years, the system scores somewhat below average, a VIQ of 88. At an assumed age of 7 years, the system scores very far below average, a VIQ of 72.

Looking at these scores in terms of percentiles makes the results clearer. The tester may choose, with constrains, which three subtests to use to compute the VIQ. In Figure 2 we show percentile results as a function of age for the (strict) standard three subtests, the lowest-scoring permissible set of three subtests, and the best-scoring permissible set of subtests. These vary greatly, because ConceptNet's subscores have much wider range than we would expect for a normal human child. Considered as a 4-year old, the system is in the 21st, 50th, and 79th percentile (worst, standard, best), while, considered as a 7-year old, the system falls below the 10th percentile using all three scoring methods.

Some Qualitative Observations

We, somewhat inadvertently, ran a comparison of two different versions of ConceptNet 4. We began our work when the most current release was the February 2010 version, which was updated in March 2012. The March 2012 release was a minor update; the number of concepts in the version of ConceptNet included in Divisi grew by about 5 percent.

The scores of the two versions of ConceptNet on the WPPSI-III were mostly extremely similar. The only really large difference was in Similarity scores. With the earlier version of ConceptNet, the strict Similarity scaled subtest score was 19 (3 standard deviations above the mean) for a 4-year old, and 11 even for a 7-year old. With the later version, the strict scaled subscale for a 4-year old was a still high, but not extraordinary, 13; however, the *relaxed* score actually went up from the old to new version. So both versions did outstandingly well in placing a correct answer somewhere among their top five answers, but evidently the best answer is not necessarily given the highest weight by ConceptNet.

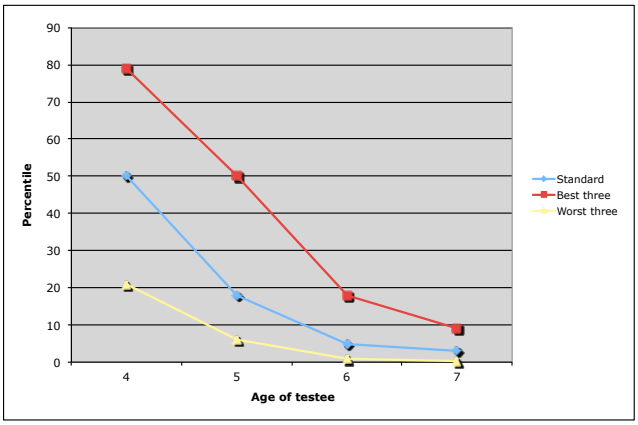

Figure 2. *WPPSI-III percentile for VIQ as a function of age, computed using the best possible, standard, and worst possible legal choice of three subtests. Strict scoring was used in all cases.*

Initially we had hypothesized Similarity items would be more difficult than either Vocabulary or Information items, because answering Similarity items requires more than mere retrieval. However, as we said, ConceptNet's results on Similarty items were consistently better than its restuls on Information items, and often better than its results on items. The high score on the Similarities subtest may reflect that abstracting categories is a particular goal of the AnalogySpace designer's use of spectral methods (Speer, Havasi, and Lieberman 2008).

Results were somewhat sensitive to whether we removed one-word concepts that were part of two-word

concepts that AnalogySpace also had. For example, in one method the translation of the Comprehension item "Why do people shake hands?" is to the single concept ['shake hand'] and in the other to the list of three concepts ['shake', 'hand', 'shake hand']. The one-concept query elicits answers of 'thank', 'flirt', and 'meet friend' with relation HasSubevent. The three-word version instead gives 'epileptic fit' HasSubevent as its top answer. Removing the one-word concepts improved performance considerably on our made-up Comprehension items, and some on the real Comprehension items. In the other direction, oddly, it hurt performance somewhat on our made-up Information items, though it made no significant difference on the WPPSI-III Information items. For example, on our made-up Information item, "Where can you find a teacher?" ['find teacher', 'find', 'teacher'] gives AtLocation 'school' as its top answer followed by AtLocation 'classroom'. But for ['find teacher'] we get AtLocation 'band' followed by AtLocation 'piano'. (The scores we report for the WPPSI-III are for the version that does remove the one-word concepts for both Information and Comprehension. We committed to that choice before running the WPPSI-III questions because it gave overall better performance on Information and Comprehension questions combined in testing on our made-up items.)

Many wrong answers are not at all like the wrong answers children would give, and seem very much to defy common sense. For example, consider the Word Reasoning item "lion" with the three clues: "This animal has a mane if it is male", "this is an animal that lives in Africa," and "this a big yellowish-brown cat." The five top answers, after all the clues, in order were: dog, farm, creature, home, and cat. Two answers, creature and cat, are in the vague neighborhood of lion. However, the other answers are crystal clear violations of common sense. Common sense should at the very least confine the answer to animals, and should also really make the simple inference that, "if the clues say it is a cat, then types of cats are the only alternatives to be considered."

Violations of common sense of this sort suggests that brittleness is hiding in the generally very ConceptNet. Large amounts of conceptual knowledge do not necessarily protect against violations.

Concluding Remarks

It is remarkable that ConceptNet can obtain an average Verbal IQ for a 4-year-old. By some definition of human intelligence, there now exists software that does have the intelligence of a young child. However, Brooks's challenge to obtain the language capabilities of 4-year olds is not met, as we considered quite different, though language-related, capabilities. We are well aware of the philosophical issues of determining whether a collection of facts subjected to a truncated SVD can be said to exhibit intelligence; here we simply speak of performance on a psychometrician's test.

We can identify several areas that limited the VIQ obtained. One is answering why questions, which make up most of the Comprehension subtest. General why questions, including both the common-sense kind discussed here and factual why questions, such as Watson answered for *Jeopardy!*, are a known difficult problem in question answering, a field at the intersection of information retrieval, natural language processing and human-computer interaction (Maybury 2004).

Other issues are related to natural language processing, missing information and incorrect information. ConceptNet does essentially no word-sense disambiguation. It combines different forms of one word into one database entry, to increase what is known about the entry. This appears to have been a deliberate choice made by ConceptNet's designers, and appears to benefit Concept Net in some situations. However, the lack of disambiguation hurts when, for example, the system's natural language processing tools convert "saw" into the base form of the verb "see," and our sample question, "What is a saw used for?" is responded by "An eye is used to see." In general, writing more powerful natural language processing tools would improve performance. The ConceptNet knowledge base does know which is the subject and which is the object in "eye UsedFor see", but the natural language processing tools that take user input to query the system do not make that distinction. Another issue is just which information the system has captured in the first place. Some information is simply missing, as we would expect. Other information is in the very large but less reliable collection of 275,000 concepts, but not in the smaller collection used by the AnalogySpace software divisi.

We speculate that improvements in those areas could improve the results to average for a five or six year old child, but that something altogether new would be needed to answer Comprehension questions (from an age-appropriate test) with the skill of a child of eight.

Still, it is remarkable that common-sense AI software has come this far.

Acknowledgements

Ohlsson was supported, in part, by Award # N00014-09-1-0125 from the Office of Naval Research (ONR), US Navy. Other authors were supported by NSF Grant CCF-0916708.

References

Brooks, R. 2008. I, Rodney Brooks, am a robot. *IEEE Spectrum* 45(6). http://cs.smith.edu/~thiebaut/research/singularity/ieee_spectrum__i,_rodney_brooks,_am_a_robot.pdf.

Davis, E. 1990. *Representations of Commonsense Knowledge*. Morgan Kaufmann Pub.

Fagin, R.; Halpern, J. Y.; Moses, Y. and Vardi, M. Y. 1995. *Reasoning About Knowledge*. The MIT Press.

Fahlman, S. 2006. Marker-passing inference in the scone knowledge-base system. *Knowledge Science, Engineering and Management*:114–126.

Gärdenfors, P. 1988. *Knowledge in Flux: Modelling the Dynamics of Epistemic States*. MIT Press.

Havasi, C.; Pustejovsky, J.; Speer, R. and Lieberman, H. 2009. Digital intuition: Applying common sense using dimensionality reduction. *Intelligent Systems, IEEE* 24(4):24–35.

Havasi, C.; Speer, R. and Alonso, J. 2007. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent Advances in Natural Language Processing*, 27–29. http://www.media.mit.edu/~jalonso/cnet3.pdf.

Lenat, D. B. 1995. Cyc: a large-scale investment in knowledge infrastructure. *Communications of the ACM* 38(11):33–38.

Maybury, M. T. 2004. Question answering: an introduction. In *New Directions in Question Answering*, 3–18.

McCarthy, J. 1959. Programs with common sense. In *Proc. Teddington Conf. on the Mechanization of Thought Processes*. http://www-formal.stanford.edu/jmc/mcc59.html.

Nilsson, N. J. 2009. *The Quest for Artificial Intelligence*. Cambridge University Press.

Ohlsson, S.; Sloan, R. H.; Turán, G.; Uber, D. and Urasky, A. 2012. An approach to evaluate AI commonsense reasoning systems. In *Twenty-Fifth International FLAIRS Conference*. http://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS12/paper/viewPDFInterstitial/4399/4828.

Perkins, D. 1995. *Outsmarting IQ: The Emerging Science of Learnable Intelligence*. Free Press.

Singh, P. 2002. The public acquisition of commonsense knowledge. In *Proc. AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*.

Speer, R.; Havasi, C. and Lieberman, H. 2008. Analogyspace: Reducing the dimensionality of common sense knowledge. In *Proceedings of AAAI*.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind* 59(236):433–460.

Wechsler, D. 2002. *WPPSI-III: Technical and Interpretative Manual*. The Psychological Corporation.