An Approach to Evaluate AI Commonsense Reasoning Systems*

Stellan Ohlsson and **Robert H. Sloan**
University of Illinois at Chicago

György Turán
University of Illinois at Chicago,
University of Szeged

Daniel Uber and **Aaron Urasky**
University of Illinois at Chicago

Abstract

We propose and give a preliminary test of a new metric for the quality of the commonsense knowledge and reasoning of large AI databases: Using the same measurement as is used for a four-year-old, namely, an IQ test for young children. We report on results obtained using test questions we wrote in the spirit of the questions of the Wechsler Preschool and Primary Scale of Intelligence, Third Edition (WPPSI-III) on the ConceptNet system, which were, on the whole, quite strong.

1 Introduction

Computer systems have long outperformed humans on numerical calculations and certain other technical tasks, and they now have human-level expertise in chess and a few other domains. However, much of what we refer to as intelligence pertains to common sense, which we define as reasoning based on specific knowledge about mundane objects, events, actions, and states of affairs.

Capturing common sense has been a central goal for Artificial Intelligence (AI) since its very beginning (McCarthy 1959). However, capturing common sense in a computer system has turned out to be difficult. 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. Systems produced by this research strategy include Cyc (Lenat and Guha 1989; Lenat 1995), Scone (Fahlman 2006), and ConceptNet/AnalogySpace (Speer, Havasi, and Lieberman 2008).

How can we evaluate claims that such systems approach human intelligence? Unlike technical or game-playing expertise, common sense reasoning has no unique outcome. Rather, common sense acts as a substrate upon which all human reasoning is based. Intuitively, common sense reasoning consists of exactly those abilities that young human children possess and AI systems often lack. That is, "The common knowledge about the world that is possessed by every schoolchild and the methods for making obvious inferences from this knowledge are called common sense" (Davis 1990). Without a generally accepted performance standard, it is impossible to evaluate claims and document progress.

We propose that tests of intelligence developed by psychometricians can serve as one type of assessment of common sense reasoning systems. Psychologists face the same problem as artificial intelligence researchers: How to measure something as broad and varied as common sense? Their solution is to generate a multitude of diverse but simple tasks, called *test items,* and then collect empirical data on the association between performance on the proposed test items and some criterion variable. This statistical solution is encapsulated in what are generally known as intelligence tests. Such tests provide a ready-made standard against which a common sense reasoner, human or machine, can be compared.

We focus on the The Wechsler Preschool and Primary Scale of Intelligence, Third Edition (WPPSI-III) test, which is a multi-dimensional IQ test designed to assess the intelligence of children of ages 2.5–7.25 years (Wechsler 2002), and is one of two widely used IQ tests for young children (the other being the Stanford-Binet). The specific questions of the WPPSI-III are proprietary (to the Psychological Corporation of America, which is a subsidiary of Pearson). However, the general nature of each subtest is public information, available from (Lichtenberger and Kaufman 2004), and various websites (e.g., Wikipedia's WPPSI-III entry and Pearson's own website).

For this preliminary work, we selected five subscales of the WPPSI-III. In order to demonstrate the plausibility of the approach we propose, we created our own sets of items, and we report our results for these questions. Our results should be informative with respect to the question of how closely current common sense reasoning systems approach the intelligence of children, and where the differences, if any, appear.

2 ConceptNet

For this proof-of-concept-work, we focus on ConceptNet (Havasi, Speer, and Alonso 2007; Speer, Havasi, and Lieberman 2008), an open-source project run by the MIT Common Sense Computing Initiative. The project has several components. The Open Mind Common Sense initiative acquired a

*The second through fifth authors were supported by NSF Grant CCF-0916708.

very large common-sense knowledge base from web users (Singh 2002). This is ConceptNet itself. AnalogySpace is a concise version of the knowledge base (Speer, Havasi, and Lieberman 2008), based on using Principal Component Analysis (PCA) on the large matrix of ConceptNet. We worked with the tools provided with the ConceptNet 4 release, which include software for working with both ConceptNet and AnalogySpace, and some limited tools for translating between them and natural language.

Both ConceptNet and AnalogySpace work with the same fixed set of about 20 relations, including IsA, HasA, UsedFor, At Location, and CapableOf; the full list is available in the documentation at `http://csc.media.mit.edu/docs/conceptnet/conceptnet4.html`. Each entry in AnalogySpace consists of two concepts and one of the relations, and a numerical strength.

3 Description of Work

The WPPSI-III consists of 14 subtests each either core, supplemental, or optional. We chose the five subtests that can contribute to the Verbal IQ Score: Vocabulary (core), Information (core), Similarities (supplemental), Word Reasoning (core), and Comprehension (supplemental). We created 18 Vocabulary items, 13 Information items, 22 Similarities items, 10 Word Reasoning items, and 12 Comprehension items.

The Five Subtests We Used

Vocabulary items ask for the meaning of a common word. For example, the testee might be asked, "What is a house?" We included a few harder challenges, such as "What is democracy?" 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.

Information items ask for the properties, kind, function, or other aspect of some everyday object, event, or process. For example, the testee might be asked, "Where can you find a penguin?" Whereas Vocabulary items request definitional or meaning information, Information items request the retrieval of factual information. However, that distinction is not sharp or precise in terms of human cognitive function.

In a Similarities item, the testee is given two concepts (words) and asked to relate them. For example, the testee might be asked, "Finish what I say. Pen and pencil are both ___." Performance on a Similarities item requires the retrieval of the two concept definitions (meanings), and finding a meaningful overlap between them.

In a Word Reasoning item, the testee must identify a concept based on one to three clues given by the tester. For example, "You can see through it," and additionally, "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 testee is asked 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, the Comprehension subtest is often described as being a test of "common sense."

Our Methodology for Querying ConceptNet

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

Vocabulary Questions We queried the following words: *Dog, Sun, Running, Happy, Crayon, Airplane, Chair, House, Baby, Shirt, Silverware, Between, Container, Democracy, Truth, Painful, Knowledge,Time*. 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.
3. For the top AnalogySpace item, 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 *dog*, the top entry from AnalogySpace we get in Step 2 relates *dog* to *pet*. In Step 3, we find that the top ConceptNet assertion for *dog* and *pet* is `<Assertion: IsA(dog, pet)[]>`. Next, in Step 4, we find the top raw assertion underlying the assertion IsA(dog, pet) is `[('dogs' IsA 'pet')`. Finally, in Step 5, ConceptNet's natural language tools translate that raw assertion into the sentence, "Dogs are a type of pet." As examples of the final answers, our top three results in order, best to worst for *dog* were: (1) dogs are a type of pet, (2) dogs can be friendly, and (3) dogs can swim. For *happy* we obtained (1) happy is when you feel good, (2) happy has joy, and (3) if you want to celebrate then you should be happy.

Information, Comprehension, and Word Reasoning 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?"

We feed the question into ConceptNet's natural language tools, which remove common stop words and return a set of ConceptNet concepts. We then create an AnalogySpace category from those concepts, which is effectively a column vector of concepts, and 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 some restrictions, such as if the first word of the input question was one of the W-question words *what*, *where,* or *why*,

then we use that word to restrict the set of relations we consider. For example, for *where* questions, we considered only the two relations AtLocation and LocatedNear. Otherwise, we use ConceptNet for question answering precisely as proposed in its documentation and tutorials.

We use essentially the same procedure for Word Reasoning, but with no special treatment of W-question words (which typically would not occur for these questions anyway).

Similarities For similarities, our inputs are two words, such as *snake* and *alligator*. For each word, we find the concept for the word and its two closest neighbors using the spreading activation function of AnalogySpace, and for each of those concepts, we find the 100 highest rated features and their scores. We create one set for each word, which could have up to 300 entries, though typically has many fewer because of common entries. We then find the features in the intersection of the two sets, and return as our answer the highest scored feature, where we determine score by adding the score from each set.

Scoring

We subjectively scored each answer using the same scales (either 0–1 or 0–1–2 depending on the subtest) as the WPPSI-III, which is the only practical thing to do for the questions we created ourselves for training and development purposes. For the WPPSI-III, there is an elaborate scoring manual that gives the score that should be assigned for many common answers, although the examiner still sometimes has to use his or her subjective judgment (Lichtenberger and Kaufman 2004).

To see how heavily the system's performance was influenced by the relative scores/weights that the system gave to different concepts, we also performed an alternate scoring where we assigned the maximum score that any of the system's five highest weighted responses would have earned. We call this score *relaxed*, and the regular score *strict*.

4 Results of Work

Quantitative Findings

We converted the raw scores into percentage of the maximum score on each subtest, as well as for the average of the five tests. With the strict scoring method, ConceptNet scored 42 to 68% on the different subtests, with a mean of 56% over the five subtests. With the relaxed scoring method, its performance rises to a range of 75 to 86%, with a mean of 79%. Thus there is a large difference in the score as a function of scoring method. Interestingly, for almost all the questions where the relaxed score was higher than the strict score, the higher-scoring answer was the system's number two or three answer, not in fourth or fifth place. Under either scoring regiment, ConceptNet is handling a significant proportion of the items correctly.

The five subtests can be ordered by the depth of processing required. Vocabulary and Information are primarily retrieval tasks, while Similarities and Word Reasoning require the system to compute the convergence of multiple retrieval processes. Finally, Comprehension requires some constructive, explanatory processing. Figure 1 shows the strict and relaxed proportional scores for the system for each subtests.

As Figure 1 shows, we see only some of the expected performance decrement as a function of item complexity: ConceptNet does quite well on Information and Word Reasoning, and least well on Comprehension. Interestingly, performance order is the same for both the strict and the relaxed scoring regimens. Surprisingly, ConceptNet does quite well on Similarities, and not so well on Vocabulary. To explain the surprises, we turn to a qualitative analysis of some of the individual items and how they are handled by ConceptNet.

Qualitative Observations

Some items elicit multiple sensible or correct answers. For example, Similarity item "canoe" and "aircraft carrier" elicited the answers *boat* (top answer), *water*, and *ocean*, which are all reasonable. However, other items elicited a mixture of sensible and non-sense answers. For example, Information item "Where does your food go when you swallow it?" yielded both *refrigeration* (top answer) and *stomach*. Given that the goal of semantic reasoning systems is to mimic common sense, answers that are clearly nonsensical are problematic.

The high 68% strict score on the Similarities subtest may reflect the fact that abstracting categories is a particular goal of the AnalogySpace's use of spectral methods. The relatively low 50% score on Vocabulary items is surprising. Of course, we may have written relatively more difficult items for this test. Vocabulary answers were also often difficult to score: Should "A dog is a pet," earn one or two points of a possible two in response to the question "What is a dog?"

We close this section with an example of a wrong answer. Consider the question, "Why do people shake hands?" from the Comprehension subtest. The main idea in the list of answers is that people shake hands in order to hold on to the other person: *to hold, to grasp,* and *to grab* were the three top answers, and numbers four and five stated that playing the harp and tying ones shoelaces are prerequisites for shaking hands. The notions of friendship, greeting, and respect were nowhere to be seen.

5 Discussion and Conclusions

ConceptNet performed well on the IQ-test type items that we used. However, this preliminary investigation is limited in several ways. First, we cannot yet assign ConceptNet an IQ score, because these preliminary results were obtained with test items of our own making, and we have no norming data for these. Second, ConceptNet's performance may have been artificially high because the same team was designing the test questions and the methodology for using ConceptNet to answer them.

A third limitation is that there is some ambiguity at the input and output ends of this enterprise. A verbal, natural language test items needs to be translated into the appropriate input for the computer system, and there are choices to be made that could affect the test score. Similarly, the interpretation of the answers is necessarily somewhat subjective.

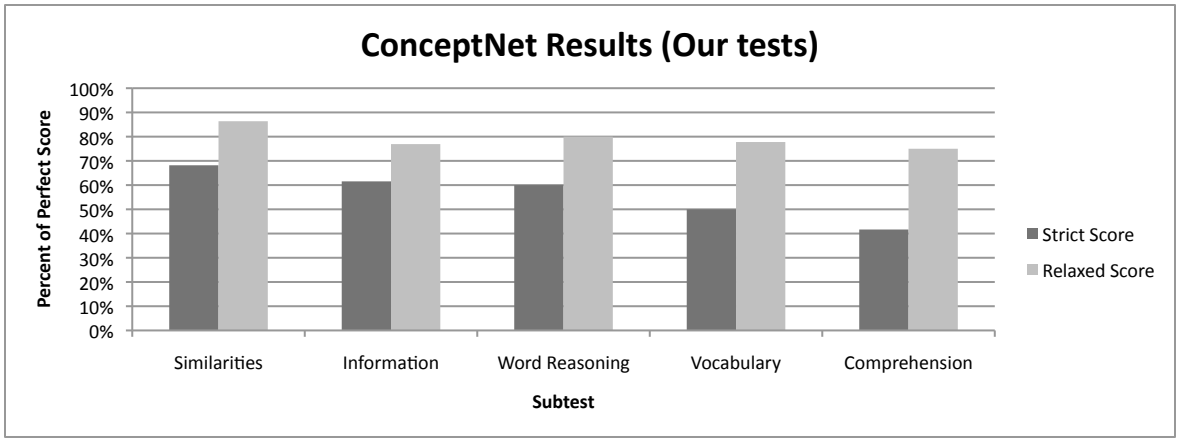

Figure 1: Results of ConceptNet on each of the five subtests for the items we made up.

A fourth limitation is that we have no baseline performance measure. For example, we do not know what level of performance could be achieved with latent semantic analysis or other statistical approaches that deliberately refrain from encoding symbolic content.

In addition to methodological limitations, our investigation uncovered some substantive puzzles. One puzzle is that the system occasionally returns answers that are more characteristic of technical expertise than of common sense, let alone the sense of a four-year old, such as the "mammal" meaning for dog, instead of "animal." The answer is correct from the point of view of biology, but it does not provide evidence for similarity between the processing of the system and the processing of a child. The response that AI systems to be useful should strive to be correct rather than to be like a child overlooks the fact that a system "performs like a four-year old" has been advanced as evidence that its processes and mechanisms are on the track towards "true" intelligence.

Another issue arises when a common sense reasoning system returns a seriously wrong answer. The goal of implementing common sense is stimulated, in part, by the wish to cure AI systems of their brittleness, the unfortunate tendency to produce reasonable and intelligent answers or actions with respect to one question, problem, or situation, but then produce an entirely nonsensical response to a seemingly similar question, problem, or situation. Rather than graceful degradation with decreasing familiarity, as in humans, AI systems often exhibit a sharp boundary between sense and nonsense. In our interactions with ConceptNet, the system's responses are mostly sensible, but there are signs that brittleness might be hiding among its links: When asked, "What is inside of tires?", the system gave the comically incorrect "Sleep MotivatedByGoal tires" as its top response. Frequent problems of this sort, even at high performance levels, would undermine the belief that common sense is a cure for brittleness.

The main conclusion of our study is that common sense semantic reasoning systems have advanced to the point at which the questions of how to assess their performance and how to evaluate claims and measure progress have become relevant. ConceptNet's performance is such that the differences and similarities between its answers and those expected of a person can meaningfully be discussed. Our methodology can be extended to other common sense reasoning systems and other tests. We have obtained the actual, normed, WPPSI-III, and we will determine a four-year-old IQ for first ConceptNet, and eventually other systems.

References

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

Fahlman, S. E. 2006. Marker-passing inference in the Scone knowledge-base system. In *Proc. First International Conference on Knowledge Science, Engineering and Management*, 114–126.

Fahlman, S. E. 2011. Using Scone's multiple-context mechanism to emulate human-like reasoning. In *2011 AAAI Fall Symp. Series*.

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.

Lenat, D. B., and Guha, R. V. 1989. *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

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

Lichtenberger, E. O., and Kaufman, A. S. 2004. *Essentials of WPPSI-III Assessment*. Wiley.

McCarthy, J. 1959. Programs with common sense. In *Proc. Teddington Conf. on the Mechanization of Thought Processes*, 756–91.

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 *Proc. AAAI Conf. on Artificial Intelligence (AAAI)*.

Wechsler, D. 2002. *WPPSI-III: Technical and interpretative manual*. The Psychological Corporation.