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

gyt@uic.edu

Aaron Urasky  
University of Illinois at Chicago

aaron.urasky@gmail.com

Abstract

Verbal tasks that have traditionally been difficult for computer systems but are easy for young children are among AI’s “grand challenges”. We present the results of testing the ConceptNet 4 system on the verbal part of the standard WPPSI-III IQ test, using simple test-answering algorithms. It is found that the system has the Verbal IQ of an average four-year-old child.

Overview and Background

The question how machines might exhibit common sense is implicit in (Turing 1950), and explicit in (McCarthy 1959). Common sense is “common knowledge about the world that is possessed by every schoolchild and the methods for making obvious inferences from this knowledge” (Davis 1990). Nilsson (Nilsson 2009) quotes Brooks’ challenges (Brooks 2008): the object-recognition capabilities of a two-year old, the language capabilities of a four-year old, the manual dexterity of a six-year-old, and the social understanding of an eight-year-old.

Capturing language capabilities and common sense in a computer system has turned out to be even harder than capturing technical expertise. Researchers have come to realize that 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).

We report on the Verbal IQ (VIQ) of ConceptNet 4 and AnalogySpace, a hybrid between symbolic and statistical systems, that uses both a semantic net and Principal Component Analysis. More information about it is available at http://csc.media.mit.edu/docs/conceptnet/. (We made a preliminary attempt to test Cyc as well but could not obtain any positive results; we wish we could measure the performance of Watson of Jeopardy! fame.)

We measure the system using a tool developed by psychometricians: the IQ test, specifically the Wechsler Preschool and Primary Scale of Intelligence, Third Edition (WPPSI-III) test, designed to assess the intelligence of young children (Wechsler 2002). Our general plan and some preliminary results on made-up questions are given in (Ohlsson et al. 2012). We report here on ConceptNet’s performance on the WPPSI-III VIQ. As required, grading was done by a psychologist (author Ohlsson). The results depend on both ConceptNet and our algorithms that use it to answer the questions.

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). These turn out to be the five subtests from which the Verbal IQ is defined.

A WPPSI-III full-scale IQ is determined by a Performance IQ (drawing, puzzle, and memory tasks) and a 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. The first three are “core,” the last two are “supplemental.” The examiner may choose any three subtests such that at least two are core.

Vocabulary items ask for word meanings, and Information items ask for simple retrieval of facts. Similarity items ask what two things (e.g., apples and oranges) have in common. Word Reasoning items ask for a word based on clues. In a Comprehension item, the task is to give an explanation in response to a why-question such as “Why do we keep ice cream in the freezer?”, going beyond retrieval. (ConceptNet answered: “UsedFor keep cold”.) The Comprehension subtest is often described as a test of “common sense” as opposed to some other cognitive ability.
Methods and Results

<table>
<thead>
<tr>
<th>Subtests included in VIQ</th>
<th>Scoring Regimen</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>Information</td>
<td>x</td>
</tr>
<tr>
<td>Word Reasoning</td>
<td>x</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>x</td>
</tr>
<tr>
<td>Similarities</td>
<td>x</td>
</tr>
<tr>
<td>Comp.</td>
<td>x</td>
</tr>
</tbody>
</table>

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

Our objective was to explore the capabilities available in ConceptNet for a non-expert user without a significant knowledge engineering effort. Thus, we wrote short Python programs to feed items into ConceptNet using its fairly rudimentary NLP tools with some added processing.

We got answers from the system with a score showing, intuitively, its degree of belief. The strict score uses the top-scored answer, and the relaxed score uses the best answer from among the top five.

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 Similarity.

Treating the testee as 4 years old, the VIQ score is average (VIQ = 100, based on scaled 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, it scores very far below average, a VIQ of 72.

The tester may choose, with constraints, 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 and the best scoring permissible set of subtests. These vary greatly, because ConceptNet’s subscores have a wider range than a typical 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.

Concluding Remarks

ConceptNet obtained an average VIQ for a 4-year-old. By some definition of human intelligence, it has the intelligence of a young child. One area for improvement is answering why questions, a known hard problem in question answering (Maybury 2004). We speculate that better algorithms could get the results to average for a five or six year old, but something new would be needed to answer Comprehension questions with the skill of a child of eight.

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

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.

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


