How Well Do Machines Perform on IQ tests: a Comparison Study on a Large-Scale Dataset

Yusen Liu[1] , **Fangyuan He**[1] , **Haodi Zhang**[2] , **Guozheng Rao**[1] , **Zhiyong Feng**[1] and **Yi Zhou**[3,4*]
[1]School of Computer Science and Technology, Tianjin University
[2]College of Computer Science and Software Engineering, Shenzhen University
[3]Shanghai Research Center for Brain Science and Brain-Inspired Intelligence/Zhangjiang Laboratory
[4]School of Natural and Computational Sciences, Massey University

Abstract

AI benchmarking becomes an increasingly important task. As suggested by many researchers, Intelligence Quotient (IQ) tests, which is widely regarded as one of the predominant benchmarks for measuring human intelligence, raises an interesting challenge for AI systems. For better solving IQ tests automatedly by machines, one needs to use, combine and advance many areas in AI including knowledge representation and reasoning, machine learning, natural language processing and image understanding. Also, automated IQ tests provides an ideal testbed for integrating symbolic and subsymbolic approaches as both are found useful here. Hence, we argue that IQ tests, although not suitable for testing machine intelligence, provides an excellent benchmark for the current development of AI research. Nevertheless, most existing IQ test datasets are not comprehensive enough for this purpose. As a result, the conclusions obtained are not representative. To address this issue, we create IQ10k, a large-scale dataset that contains more than 10,000 IQ test questions. We also conduct a comparison study on IQ10k with a number of state-of-the-art approaches.

1 Introduction

Due to the rapid development and massive volume of AI research, AI benchmarking becomes an increasingly important task [Plan, 2016]. Meaningful AI benchmarks, such as ImageNet [Russakovsky *et al.*, 2015] and RoboCup [Kitano *et al.*, 1997], not only provide a standard testbed for comparing different AI approaches, but also significantly promote and stimulate AI research.

It is well known that Intelligence Quotient (IQ) test is widely recognized as one of the predominant benchmarks for measuring human intelligence [Rowe *et al.*, 2012; Dowe and Hernández-Orallo, 2014]. A natural problem arises whether it can serve as a meaningful AI benchmark as well. Indeed, this has been seriously discussed in the literature [Selmer Bringsjord, 2003; Detterman, 2011; Hernández-Orallo *et al.*, 2016]. In this paper, we argue that IQ tests, although not suitable for testing machine intelligence, provides an excellent benchmark for the current development of AI research, mainly for the following two reasons. Firstly, for better solving IQ tests automatedly by machines, one needs to use, combine and advance many areas in AI including knowledge representation and reasoning, machine learning, natural language processing and image understanding. Secondly, as both symbolic approaches and subsymbolic approaches are proven to be useful in automated IQ tests, it provides an ideal testbed for integrating these two critical AI research lines.

There are a few IQ test datasets in the literature [Lynn and Vanhanen, 2009; Pietschnig and Voracek, 2015; Wang *et al.*, 2016]. However, most of them are not comprehensive enough. Firstly, in terms of volume, many of them only contain dozens or up to a few hundreds of questions. Secondly, in terms of variety, most of them only contain a single category, e.g., number sequence. As a consequence, the conclusions obtained from existing IQ test datasets, although valuable, are not representative. Also, there are a number of datasets that are highly related to IQ test questions as well. For instance, the Online Encyclopedia of Integer Sequences (OEIS) [Sloane and others, 2007] contains over a quarter-million number of math sequences, Also, Siebers et al. [2012] used a random generator to construct many integer sequences. However, these datasets do not directly target on IQ tests, thus are essentially different.

To address this issue, we construct IQ10k, a large-scale dataset that contains more than 10,000 IQ test questions. For this purpose, we manually collect IQ test questions together with their answers and solution hints from IQ test books, websites and other resources. Depicted in Table 1, we group the questions into four major categories, namely Verbal, Sequence, Diagram and Other. Among them, the Verbal category is the biggest, which contains 4503 questions. The Sequence category follows right after. Diagram contains some questions to find the most appropriate diagram given a number of others. The rest are grouped into the "Other" category that mainly contains some math, logic or commonsense reasoning problems. As we will explain later in the paper, these categories can further be divided into some subcategories. For each question, we record its question, its answer, its category and its solution hints if any.

We conduct a comparison study to see how existing ap-

*corresponding author

proaches perform on IQ10k. In this paper, we focus on the Verbal category and the Sequence category. For Sequence, we consider four representative approaches, namely the OEIS website,[1] Mathematica,[2] BathSeq - a semi-analytical approach [Siebers and Schmid, 2012], and an artificial neural network based approach [Ragni and Klein, 2011]. Verbal itself contains many subcategories such as analogy and classification. So far, there is no work that can solve them all. Hence, we select two of its most representative subcategories, namely Analogy and Classification. In fact, there are only a few approaches directly targeting on solving IQ Verbal questions. Nevertheless, the well known word similarity methods can be straightforwardly applied here. Hence, we also apply some representative word embedding approaches, including Word2Vec [Mikolov *et al.*, 2013], GloVe [Pennington *et al.*, 2014] and ConceptNet Numberbatch [Speer *et al.*, 2017], to solving Verbal questions.

Our experimental results show that existing approaches, although they perform much better than random guess, are still worse than human being on average. Moreover, normally these approaches can only deal with a single type of questions. Hence, we argue that automated IQ test provides an interesting and suitable benchmark for the current development of AI research. Evident from other AI areas including machine learning, computer vision, speech recognition and natural language processing, large-scale datasets such as ImageNet [Russakovsky *et al.*, 2015], not only provided better testbeds for justifying AI systems, but also significantly stimulated research in the corresponding areas. We hope that, with IQ10k, more important AI technologies can be developed and tested.

2 Related Work

As discussed in the introduction section, AI benchmarking and its associated meaningful large-scale datasets become an increasingly important task for the current AI research and development. Among all, IQ tests have been suggested by many researchers to be a meaningful task for this purpose [Selmer Bringsjord, 2003; Bringsjord, 2011; Detterman, 2011; Hernández-Orallo *et al.*, 2016]. Detterman [2011] strongly supported this idea and raised a challenge to the AI community for constructing a unique battery of tests for this purpose. Hernández-Orallo et al. [2016] observed that computer models addressing intelligence tests have different purposes and applications, not only (a) to advance AI by the use of challenging problems from a psychometric AI perspective and (b) to use them for the evaluation of AI systems, but also (c) to better understand intelligence tests and what they measure, and (d) to better understand what human intelligence is.

[1]https://oeis.org/

[2]https://www.wolfram.com/mathematica/

All	Verbal	Sequence	Diagram	Other
10007	4503	2562	1205	1737

Table 1: An overview of IQ10k

There are also some debates in the literature. Dowe et al. [2003] showed that some IQ test questions are too easy so that they can be solved by a simple program. Nevertheless, as we will show later in this paper, existing approaches still perform worse than human for solving IQ test questions in a large-scale dataset. Hernandez-Orallo [2017] pointed out that IQ tests are specifically designed for testing general human intelligence. It is questionable whether they are suitable for testing general machine intelligence as well. We agree that IQ tests cannot serve as a standard test for measuring general machine intelligence, unlike what they do for human being[3].

Nevertheless, we argue that IQ test provides an ideal benchmark for the current development of AI research for the following reasons. Firstly, IQ tests cover many different categories, e.g., verbal and logic questions, so that it can test many aspects of AI technologies, including knowledge representation and reasoning, machine learning, natural language processing and image processing. For better solving IQ tests, one needs to use, combine and advance techniques in these AI subareas. Secondly, evident from our experiments, both symbolic approaches and subsymbolic approaches are quite useful. Hence, automated IQ tests provides an ideal testbed for integrating these two critical AI research lines. Thirdly, IQ tests mainly focus on discovering underlying patterns and principles by given samples and data in various areas including logic and verbal, which is one of the major aspects of intelligence. Last but not least, although IQ test questions cover many AI aspects and are challenging to AI systems, they are relatively simple on the presentations of questions and solutions. This perfectly fits into the current development of AI research that fails to deeply understand natural languages and imagines.

There is a long history in the AI literature for automatedly solving IQ test questions or some similar problems [Forbus *et al.*, 2005; Ragni and Klein, 2011; Siebers and Schmid, 2012; Turney, 2012; Hofmann *et al.*, 2014; Wang *et al.*, 2016; Bayoudh *et al.*, 2012; Ohlsson *et al.*, 2017]. For space reasons, we omit some references, many of which can be found in an excellent survey conducted by Hernandez et al. [2016].

There are mainly four categories of questions, namely

1. Verbal: given some words under a context, e.g., analogy, to find the correct word,
2. Sequence: given a sequence of numbers, to find the next or a missing number,
3. Diagram: given some diagrams in an order, to find the most appropriate diagram,
4. Other: given a math, logic or commonsense problem, to solve it by reasoning.

Among all, Verbal and Sequence have attracted most AI researchers and have achieved relatively good results. In contrast, Diagram and Other are less studied. We rule out some type of questions, e.g. memory and response time, as they can be easily solved by machines.

The Verbal category can be further divided into many subcategories including analogy, classification, synonym and

[3]This is even debatable in the psychometrics community [Rowe *et al.*, 2012; Dowe and Hernández-Orallo, 2014].

Analogy	Classification	Antonym	Word Composition
Mohair is to wool as shantung is to: <A>silk <B>cotton <C>linen <D>nylon <E>fabric Answer: A Category: verbal-single analogy Hint: Function material	Which is the odd one out <A>heptagon <B>triangle <C>hexagon <D>cube <E>pentagon Answer: D Category: verbal-classification Hint: cube is three-dimensional figure, the rest are all two-dimensional figures	Which word is most opposite to narcissistic? <A>conceited <B>egotistic <C>self-conscious <D>self-centred Answer: C Category: verbal-antonym Hint: narcissistic means self-centred and the closest opposite is someone who is self-conscious	Night long boat house? What comes next? <A>calm <B>hold <C>panic <D>wind <E>post Answer: hold Category: verbal-word composition Hint: to form compound word: nightlong, longboat, boathouse, household

Figure 1: Sample Verbal questions

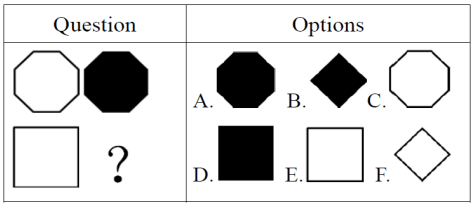

Figure 2: An example of Raven's progressive matrices

antonym, and other types such as pronunciation. Previous studies are mainly focused on the first three. For instance, Wang et al. [2016] proposed a deep learning based approach on these types of questions and claimed that it outperforms human being on average. Nevertheless, their test dataset only contains 223 questions altogether for all four types. Also, a number of work have been done on Scholastic Assessment Test (SAT) analogy questions [Turney, 2012; Nakov and Hearst, 2008; Bayoudh *et al.*, 2012]. Although similar and related, SAT analogy questions are essentially different from IQ analogy questions.

Sequence questions are also considered in the literature. Holzman et al. [1983] defined a set of characteristics of number sequences. The well known software Mathematica itself provides a prediction for number sequences as well. Other than these two, a number of different approaches are proposed. For instance, Ragni et al. [2011] proposed a neural network based method. Siebers et al. [2012] considered a semi-analytically approach. Strannegård et al. [2013] implemented a number of solvers including ASolver and SeqSolver. Hofmann et al. [2014] developed a system called IGOR2 that uses inductive programming as the foundation. Nevertheless, as we will show later, number sequence based on mathematical functions, e.g., polynomials, are essentially different from IQ Sequence questions.

In contrast, Diagram questions and Other questions are less studied. Previous studies on diagram questions main focused on three subcategories, namely geometric-analogy, odd-one-out and Raven's Progressive Matrices (RPM). A number of approaches [Forbus *et al.*, 2005; Ragni and Neubert, 2014; Lovett and Forbus, 2017] have been proposed for one of the subcategories. The "Other" category include some math, logic and commonsense reasoning problem, which have been largely ignored except a few attempts [Ohlsson *et al.*, 2017] in the literature.

Although a number of approaches have been proposed to solve IQ questions in the literature, there are several critical drawbacks. Firstly, the datasets used in previous works are either too small [Bayoudh *et al.*, 2012; Ohlsson *et al.*, 2017; Strannegård *et al.*, 2013; Wang *et al.*, 2016], or essentially different from IQ test questions [Turney, 2012; Nakov and Hearst, 2008; Bayoudh *et al.*, 2012]. Secondly, there is no comprehensive systematic comparison study so far. Last but not least, almost all approaches only focus on one specific (sub)category. Hence, we believe that, in order to promote IQ tests as a meaningful AI benchmark, there is a need to construct a large-scale IQ test dataset, and to conduct a thorough comparison study based on it.

3 Dataset

One of the most critical issues in the current research on automated solving IQ tests is that the datasets used are relatively small. As a consequence, the conclusions drawn based on which, although valuable, may not be representative. In order to address this issue, we construct a large-scale IQ test dataset. We name it as IQ10k since it contains more than 10,000 IQ test questions, more than 40 times larger than all previous IQ datasets. Followings are the main properties of IQ10K.

3.1 Scale

As shown in Table 1, so far, IQ10k contains 10,007 number of questions. They are grouped into four main categories, namely Verbal, Sequence, Diagram and Other. We use XML to unify the format of these questions.

Questions in IQ10k are manually collected from multiple resources, including IQ test books and related websites. Different from other related questions such as SAT analogy, all questions that we collected are genuine IQ test questions. The construction process consists of four steps: 1. proposal, 2. three-round review, 3. approval and 4. formalization, for each single question in the dataset.

3.2 Hierarchy

The four main categories can be further divided into subcategories. Each subcategory contains a number of different types

Math	Logic	Common sense
If you have a cube which is 5m*5m*5m, what is the cubic metres this container would hold ? Answer: 125m^3 Category: other-math Hint: 5m*5m*5m=125m^3	If all Bloops are Razzies and all Razzies are Lazzies, then all Bloops are definitely Lazzies? <A>True <B>False Answer: A Category: other-logic Hint: Reasoning	Leap years have 1 day fewer than standard years? <A>Fact <B>Fiction Answer: B Category: other-commonsense Hint: They have 1 day MORE, 366 days, the extra day being February 29th

Figure 3: Other questions

of IQ questions. For instance, the Analogy subcategory in the Verbal category contains questions to examine a tester's ability to define relationships between words as well as the tester's vocabulary.

The Verbal category contains some subcategories including analogy, classification, synonym and antonym, and other types such as pronunciation. Some sample Verbal questions are depicted in Figure 1.

3.3 Diversity

Like IQ tests are designed to examine one's comprehensive intelligence quotient from different aspects, as a testbed for machine intelligence, questions in IQ10K from multiple resources cover varieties of domains. Crowd workers collect questions and propose them first, and then after three rounds of reviews, we decide whether to accept them. Sequence questions in IQ tests are also different from similar math questions in other datasets [Sloane and others, 2007; Siebers and Schmid, 2012]. Firstly, in most cases, IQ Sequence questions only use simple patterns, e.g., Fibonacci, modulo and square. It is very rare that the hint is a polynomial like $3n^4 + 2n^2 - 18n + 7$. Instead, IQ Sequence questions are more concerned with combinations of those simple patterns. Secondly, IQ Sequence questions only contain a relatively small number (normally 4-8) of items. Last but not least, some IQ Sequence questions may have decimal points and fractions, but they should normally be interpreted from a structure point of view rather than a number point of view. In contrast, many other math sequence datasets mainly consider integer sequences. In IQ10k, Diagram questions also contain some subcategories, mostly from Raven's Progressive Matrices, geometrical problems and the odd-one-out problems. Figure 2 illustrates an example of Raven's progressive matrices. We store pictures locally and use their references to present questions and options. Finally, Figure 3 depicts the "Other" category, which contains a large number of different kinds of problems. Normally, prior knowledge are needed in order to solve these problems. Sometimes natural language understanding and common sense reasoning are also needed.

We believe that experimental results on a large-scale IQ test dataset such as IQ10k are much more convincing. We argue that IQ10k can serve as a challenging benchmark for the current development of AI research. More importantly, we hope that, similar to ImageNet, IQ10k can stimulate and promote AI research along this research line.

4 Experiments

As discussed previously, existing IQ test datasets are either too small or essentially different. As a result, the conclusions drawn from these datasets, although valuable, might not be representative. Hence, a natural question arises: on a large-scale IQ test dataset such as IQ10k, how well do existing approaches perform? In order to answer this question, we did a comparison study and report the experimental results in this section. Since there are a considerable number of approaches, we are only able to select some representative ones. Our selection criteria are twofold, namely how good they perform on existing datasets and how representative their methods are. Most existing approaches only target on a single (sub)category. Hence, we split our experiments category-by-category as well. In this paper, we focus on Verbal and Sequence. First, Verbal and Sequence are the two largest categories. Second, as far as we have checked, Verbal and Sequence have attracted more attention in the literature. Third, Verbal and Sequence seem easier so that they are better to serve as a starting point. Last but not least, space limit is also one of our concerns.

We also compare the performance of existing approaches with a random guess solver and human being. We invited 25 volunteers participating in our experimental studies. Each volunteer did 55 questions including both Verbal and Sequence questions randomly generated from our test questions in IQ10k. The volunteers completed 1375 questions altogether. Nevertheless, due to resource limit, we are not able to conduct massive experiments on human being.

For Verbal, since it contains many subcategories and many approaches only deal with one of them, we have to split them based on subcategories. For this purpose, we select analogy (including both labelled analogy and unlabelled analogy) and classification since they are of the most representative and attractive in the literature. Word similarity is a well investigated field in natural language processing. Although not explicitly claimed, many approaches in word similarity, e.g., word embedding, can be directly applied to solving IQ Verbal questions, for instance, by using the method proposed in RK. Hence, we also consider some representative word embedding approaches, including Word2Vec , GloVe, Retrofitting [Jauhar *et al.*, 2015] and ConceptNet. Also, there are some work directly targeting on SAT analogy questions. We consider Dual-Space, a representative one along this research direction.

For IQ Verbal analogy, after we obtain the words represen-

	Labelled analogy					Unlabelled analogy			Both analogy
option	3	4	5	6	total	4	5	total	total
Random guess	33.33%	25.00%	20.00%	16.67%	23.75%	25.00%	20.00%	22.50%	23.75%
Human being	63.16%	62.96%	62.26 %	59.57%	62.04%	35.71%	38.16%	37.75%	49.79%
Word2Vec	46.87%	38.89%	34.31%	41.02%	38.33%	22.22%	28.62%	28.32%	31.14%
GloVe	51.61%	38.89%	29.70%	33.33%	35.56%	25.92%	24.68%	24.74%	27.78%
Retrofitting	**55.56%**	43.75%	32.61%	29.17%	38.33%	**35.71%**	21.84%	22.73%	28.24%
ConceptNet	50.00%	**49.12%**	**48.04%**	**41.03%**	**47.39%**	29.63%	**48.15%**	**47.31%**	**47.34%**
Dual Space	31.25%	36.36%	31.71%	11.11%	29.73%	34.78%	34.98%	34.96%	33.65%

Table 2: A comparison study on IQ Verbal analogy questions

tations of Word2Vec, GloVe, Retrofitting and ConceptNet, we use the metric with weight constraint proposed by Speer et al. [Speer *et al.*, 2017] to solve the single and duoble analogy questions. For a question of form like 'a is to b as c is to d', according to the formula (1), the candidate word with the highest score is selected as the answer.

$$\begin{aligned} s = a \cdot b + c \cdot d + w_1 (d - c) \cdot (b - a) \\ + w_2 (d - b) \cdot (c - a) \end{aligned} \quad (1)$$

Given that Dual Space sets some rules for the analogy problem, we still use the metrics in the original paper. For a question of form 'a is to b as c is to d', a and b share the same domain, as well as c and d. a and c share the same function, as well as b and d. Thus there are two spaces in the model, domain space and function space, for measuring domain and function similarity, denoted by sim_d and sim_f respectively. Now we define the similarity sim_r as

$$sim_r (a, b, c, d) = \begin{cases} sim_1(a,b,c,d) & if\ sim_2(a,b,c,d) \geq sim_3(a,b,c,d) \\ 0 & otherwise \end{cases} \quad (2)$$

where sim_i for $i = 1, 2, 3$ is defined as

$$\begin{aligned} sim_1 (a, b, c, d) &= geo (sim_f (a, c), sim_f(b, d)) \\ sim_2 (a, b, c, d) &= geo (sim_d (a, b), sim_d(c, d)) \\ sim_3 (a, b, c, d) &= geo (sim_d (a, d), sim_d(c, b)) \end{aligned}$$

and geo is geo function:

$$geo (x_1, x_2, ..., x_n) = \begin{cases} (x_1 x_2 ... x_n)^{1/n} & if\ x_i > 0\ for\ all\ i=1,...,n \\ 0 & otherwise \end{cases}$$

For IQ Verbal classification, in order to select the different one, we first do a combination to all the candidate words in pairs. for instance, 'a, b, c, d' is the candidate word, after the combination we will get six word pairs, 'ab, ac, ad, bc, bd, cd', then we compute the cosine similarity of each pair. Next, we sort and select the minimum n scores, n is the number of candidate word minus one. Here n is 3. Suppose the word pairs with minimum scores are ab, bc, cd. We choose word a and d as my candidate answers, and then see whether a and b appear in the following word pair bc, if either of them appears, it will be the predicted answer, here b is the predicted answer. Otherwise, going on to the next word pair.

For IQ Verbal analogy, we select 927 questions from IQ10k, including 287 labelled analogy and 640 unlabelled analogy respectively. We divide them by the number of candidate choices, ranging from three to six for labelled analogy and four to five for unlabelled analogy. Table 2 depicts the performance of the representative approaches, in comparison with a random guess solver and human being. It can be observed that although almost all approaches are much better than the random guess, they are worse than human performance in general. Among them, ConceptNet [Speer *et al.*, 2017] outperforms the rest approaches and achieves a relatively high score that is closer to human performance. It can also be observed that, the more choices the questions have, the more difficult they are. It is worth mentioning that Dual-Space does not perform quite well on IQ10k, in comparison with its performance on SAT analogy questions. Perhaps one of the most important reasons is that IQ Verbal analogy questions are essentially different from SAT analogy questions, as explained earlier.

	Classification		
option	4	5	total
Human being	79.87%	49.45%	68.57%
Random guess	25.00 %	20.00 %	22.25%
ConceptNet	**77.14%**	**40.00%**	**62.61%**
Word2Vec	66.06%	32.85%	53.35%
GloVe	61.61%	31.75%	50.45%
Retrofitting	51.03%	22.67%	43.12%

Table 3: A comparison study on classification questions

For IQ Verbal classification, we select 358 questions from IQ10k. We also divide them by the number of choices, ranging from four to five. We do not consider Dual-Space here as it is only for analogy questions. Table 3 depicts the experimental results. Similar to analogy, these approaches are much better than random guess but worse than human being. Once again, ConceptNet is the winner. Interestingly, Word2Vec follows right after.

For Sequence, we select 2,000 questions from IQ10k. We rule out the rest mainly because many of them have decimals and fractions so that they are not compatible with many existing methods. In the 2,000 questions, we pick up 1,500 questions as the training set and the rest 500 as the test set. These 500 questions belong to 5 different types according to their solution hints, namely "Linear" for linear functions, "Power" for power functions, "Fibo" for Fibonacci se-

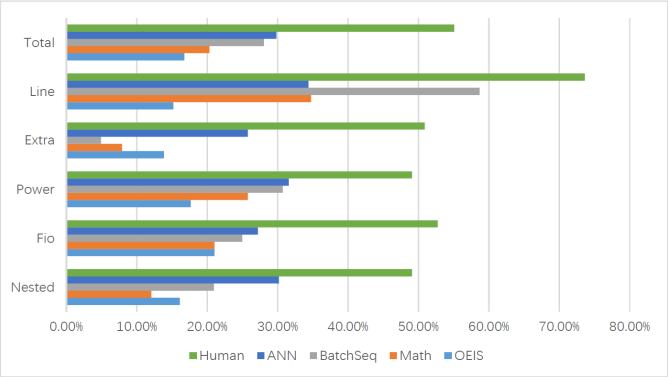

Figure 4: A comparison study on Sequence questions

quences, "Nested" for the combinations of the above and "Extra" for the rest questions that can hardly be identified as a certain type. Each type contains 100 questions.

We mainly consider the following approaches, namely OEIS - the Online Encyclopedia of Integer Sequences [Sloane and others, 2007], the well known software Mathematica from Wolfram Research [Buchberger, 1993], a semi-analytic method called BatchSeq by Siebers et al. [2012] and an artificial neural network (ANN) based method [Ragni and Klein, 2011]. We also compare the performance of these approaches with the average performance of human being. Nevertheless, we do not consider random guess as many of the Sequence questions are not multiple-choice.

Figure 4 shows the experiment results on Sequence questions. Interestingly and surprisingly, different from on Verbal questions, existing methods perform much worse than human being in general on Sequence questions. On average, human can correctly answer 55.09% of the questions while the best automated approach, namely ANN, can only solve 29.80% so far. In terms of different types, all approaches including human being perform better in simple types such as "Linear". Automated approaches perform quite bad on the "Extra" type that does not belong to any other types. Interestingly, the ANN approach's performance on these types does no show a significant difference. It is worth mentioning that the performance of existing approaches on Sequence questions in IQ10k are much worse than those on other math Sequence question datasets, e.g., OEIS. This, again, shows that IQ sequence questions are essentially different from math sequence questions.

From our experimental analysis, we conclude that existing methods perform much better than random guess, but still worse than average human being. Moreover, most of these methods only deal with a single (sub-) category. It is interesting to see that ConceptNet, that utilizes both crowdsourcing knowledge and machine learning techniques, performs very close to human being on verbal questions, including both analogy and classification. This further justifies that it is an effective approach for word embedding. Indeed, this is also one of our main motivations for constructing IQ10k. Beyond this, we hope a large-scale dataset such as IQ10k can serve as a testbed for measuring the effectiveness of more approaches and fostering new ones.

5 Conclusion

In this paper, we reported IQ10k, a new large-scale dataset that contains more than 10,000 IQ test questions. We also did a comparison study on IQ10k. The experiment results showed that existing approaches, although only targeting on one specific (sub-) category, still perform worse than human being in general. We argue that IQ test provides an interesting and meaningful benchmark for the current development of AI research. We hope that, with IQ10k, more important AI technologies can be developed and justified.

Based on IQ10k, we have established an open automated IQ test platform [4], on which AI programs can test their performance on IQ tests.

Acknowledgements

We would like to thank some related researchers for generously sharing their datasets and source codes. We also thank the anonymous reviewers of this paper and its previous versions for their valuable comments. We sincerely appreciate the student volunteers for contributing some of the IQ test questions, establishing the automated IQ test platform and participating on the IQ test experiments. This work is supported by National Natural Science Foundation of China (Grant No. NSFC-61806132), Tencent Rhino-Bird Open Fund and the Priming Research Fund in Shenzhen University.

References

[Bayoudh *et al.*, 2012] Meriam Bayoudh, Henri Prade, and Gilles Richard. Evaluation of Analogical Proportions through Kolmogorov Complexity. *Knowledge-Based Systems*, 29(3):20–30, 2012.

[Bringsjord, 2011] Selmer Bringsjord. Psychometric artificial intelligence. *Journal of Experimental and Theoretical Artificial Intelligence*, 23(3):271–277, 2011.

[Buchberger, 1993] Bruno Buchberger. Mathematica: A System for Doing Mathematics by Computer? *International Symposium on Design and Implementation of Symbolic Computation Systems*, pages 1–1, 1993.

[Detterman, 2011] Douglas K Detterman. A Challenge to Watson. *Intelligence*, 2(39):77–78, 2011.

[Dowe and Hernández-Orallo, 2014] David L Dowe and José Hernández-Orallo. How universal can an intelligence test be? *Adaptive Behavior - Animals, Animats, Software Agents, Robots and Adaptive Systems*, 22(1):51–69, February 2014.

[Forbus *et al.*, 2005] Kenneth D Forbus, Andrew Lovett, Emmett Tomai, and Jeffrey Usher. A Structure Mapping Model for Solving Geometric Analogy Problems. *Proceedings of the Cognitive Science Society*, 27(27), 2005.

[Hernández-Orallo *et al.*, 2016] José Hernández-Orallo, Fernando Martínez-Plumed, Ute Schmid, Michael Siebers, and David L Dowe. Computer models solving intelligence

[4] http://timmurphy.org/2009/07/22/line-spacing-in-latex-documents/

test problems: Progress and implications. *Artificial Intelligence*, 230:74–107, 2016.

[Hernández-Orallo, 2017] José Hernández-Orallo. Evaluation in Artificial Intelligence: from Task-oriented to Ability-oriented Measurement. *Artificial Intelligence Review*, 48(3):397–447, 2017.

[Hofmann *et al.*, 2014] Jacqueline Hofmann, Emanuel Kitzelmann, and Ute Schmid. Applying Inductive Program Synthesis to Induction of Number Series a Case Study with IGOR2. In *KI'2014*, pages 25–36, 2014.

[Holzman *et al.*, 1983] Thomas G Holzman, James W Pellegrino, and Robert Glaser. Cognitive Variables in Series Completion. *Journal of Educational Psychology*, 75(4):603, 1983.

[Jauhar *et al.*, 2015] Sujay Kumar Jauhar, Chris Dyer, and Eduard H Hovy. Ontologically Grounded Multi-sense Representation Learning for Semantic Vector Space Models. In *HLT-NAACL*, pages 683–693, 2015.

[Kitano *et al.*, 1997] Hiroaki Kitano, Minoru Asada, Yasuo Kuniyoshi, Itsuki Noda, and Eiichi Osawa. Robocup: The robot world cup initiative. In *Proceedings of the first international conference on Autonomous agents*, pages 340–347. ACM, 1997.

[Lovett and Forbus, 2017] Andrew Lovett and Kenneth Forbus. Modeling visual problem solving as analogical reasoning. *Psychological review*, 124(1):60, 2017.

[Lynn and Vanhanen, 2009] Richard Lynn and Tatu Vanhanen. Intelligence and the Wealth and Poverty of Nations, 2009.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[Nakov and Hearst, 2008] Preslav Nakov and Marti A. Hearst. Solving Relational Similarity Problems Using the Web as a Corpus. In *Proceedings of the Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, Usa*, pages 452–460, 2008.

[Ohlsson *et al.*, 2017] Stellan Ohlsson, Robert H Sloan, György Turán, and Aaron Urasky. Measuring an Artificial Intelligence System's Performance on a Verbal IQ test for Young Children. *Journal of Experimental & Theoretical Artificial Intelligence*, 29(4):679–693, 2017.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global Vectors for Word Representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[Pietschnig and Voracek, 2015] Jakob Pietschnig and Martin Voracek. One century of global IQ gains: A Formal Meta-analysis of the Flynn Effect (1909–2013). *Perspectives on Psychological Science*, 10(3):282–306, 2015.

[Plan, 2016] Strategic Plan. The National Artificial Intelligence Research and Development Strategic Plan. 2016.

[Ragni and Klein, 2011] Marco Ragni and Andreas Klein. Predicting Numbers: an AI Approach to Solving Number Series. *KI 2011: Advances in Artificial Intelligence*, pages 255–259, 2011.

[Ragni and Neubert, 2014] Marco Ragni and Stefanie Neubert. Analyzing Raven's Intelligence Test: Cognitive Model, Demand, and Complexity. In *Computational Approaches to Analogical Reasoning: Current Trends*, pages 351–370. Springer, 2014.

[Rowe *et al.*, 2012] Ellen W Rowe, Cristin Miller, Lauren A Ebenstein, and Dawna F Thompson. Cognitive predictors of reading and math achievement among gifted referrals. *School psychology quarterly : the official journal of the Division of School Psychology, American Psychological Association*, 27:144–53, 09 2012.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[Sanghi and Dowe, 2003] Pritika Sanghi and David L Dowe. A Computer Program Capable of Passing IQ Tests. In *4th Intl. Conf. on Cognitive Science (ICCS'03), Sydney*, pages 570–575, 2003.

[Selmer Bringsjord, 2003] Bettina Schimanski Selmer Bringsjord. What is Artificial Intelligence? Psychometric AI as an Answer. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, pages 887–893, 2003.

[Siebers and Schmid, 2012] Michael Siebers and Ute Schmid. Semi-analytic Natural Number Series Induction. In *KI*, pages 249–252, 2012.

[Sloane and others, 2007] Neil JA Sloane et al. The Online Encyclopedia of Integer Sequences. In *Symposium on Towards Mechanized Mathematical Assistants: International Conference*, pages 130–130, 2007.

[Speer *et al.*, 2017] Robert Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI*, pages 4444–4451, 2017.

[Strannegård *et al.*, 2013] Claes Strannegård, Abdul Rahim Nizamani, Anders Sjöberg, and Fredrik Engström. Bounded Kolmogorov Complexity Based on Cognitive Models. In *International Conference on Artificial General Intelligence*, pages 130–139, 2013.

[Turney, 2012] Peter D Turney. Domain and Function: A Dual-Space Model of Semantic Relations and Compositions. *Journal of Artificial Intelligence Research*, 44:533–585, 2012.

[Wang *et al.*, 2016] Huazheng Wang, Fei Tian, Bin Gao, Jiang Bian, and Tie-Yan Liu. Solving Verbal Questions in IQ Test by Knowledge-Powered Word Embedding. *EMNLP'16*, pages 541–550, 2016.