Toward a Comprehension Challenge, Using Crowdsourcing as a Tool

Praveen Paritosh, Gary Marcus

■ *Human readers comprehend vastly more, and in vastly different ways, than any existing comprehension test would suggest. An ideal comprehension test for a story should cover the full range of questions and answers that humans would expect other humans to reasonably learn or infer from a given story. As a step toward these goals we propose a novel test, the crowdsourced comprehension challenge (C^3), which is constructed by repeated runs of a three-person game, the Iterative Crowdsourced Comprehension Game (ICCG). ICCG uses structured crowdsourcing to comprehensively generate relevant questions and supported answers for arbitrary stories, whether fiction or nonfiction, presented across a variety of media such as videos, podcasts, and still images.*

Artificial Intelligence (AI) has made enormous advances, yet in many ways remains superficial. While the AI scientific community had hoped that by 2015 machines would be able to read and comprehend language, current models are typically superficial, capable of understanding sentences in limited domains (such as extracting movie times and restaurant locations from text) but without the sort of wide-coverage comprehension that we expect of any teenager.

Comprehension itself extends beyond the written word; most adults and children can comprehend a variety of narratives, both fiction and nonfiction, presented in a wide variety of formats, such as movies, television and radio programs, written stories, YouTube videos, still images, and cartoons. They can readily answer questions about characters, setting, motivation, and so on. No current test directly investigates such a variety of questions or media. The closest thing that one might find are tests like the comprehension questions in a verbal SAT, which only assess reading (video and other formats are excluded) and tend to emphasize tricky questions designed to discriminate between strong and weak human readers. Basic questions that would be obvious to most humans — but perhaps not to a machine — are excluded.

Yet is is hard to imagine an adequate general AI that could not comprehend with at least the same sophistication and breadth as an average human being, and easy to imagine that progress in building machines with deeper comprehension could radically alter the state of the art. Machines that could comprehend with the sophistication and breadth of humans could, for instance, learn vastly more than current systems from unstructured texts such as Wikipedia and the daily news.

How might one begin to test broad-coverage comprehension in a machine?

 ISSN 0738-4602

In principle, the classic Turing test might be one way to assess the capacity of a computer to comprehend a complex discourse, such as a narrative. In practice, the Turing test has proved to be highly gameable, especially as implemented in events such as the Loebner competitions, in which the tests are too short (a few minutes) to allow any depth (Shieber 1994; Saygin, Cicekli, and Akman 2003). Furthermore, empirical experimentation has revealed that the best way to "win" the Turing test is to evade most questions, answering with jokes and diversionary tactics. This winds up teaching us little about the capacity of machines to comprehend narratives, fictional or otherwise.

As part of the Turing Championships, we (building on Marcus [2014]) would like to see a richer test of comprehension, one that is less easily gamed, and one that probes more deeply into the capacity of machines to understand materials that might be read or otherwise perceived.

We envision that such a challenge might be structured into separate tracks for audio, video, still images, images with captions, and so forth, including both fiction and nonfiction. But how might one generate the large number of questions that provide the requisite breadth and depth? Li et al. (forthcoming) suggest one strategy, focused on generating "journalist-style" questions (who, what, when, where, why) for still images.[1] Poggio and Meyers (2016) and Zitnick et al. (2016) suggest approaches aimed at testing question answering from still images. Here, we suggest a more general procedure, suitable for a variety of media and a broad range of questions, using crowdsourcing as the primary engine.

In the remainder of this article we briefly examine what comprehension consists of, discuss some existing approaches to assessing it, present desiderata for a comprehension challenge, and then turn toward crowdsourcing and how it can help define a meaningful comprehension challenge.

What Is Human Comprehension?

Human comprehension entails identifying the meaning of a text as a connected whole, beyond a series of individual words and sentences (Kintsch and van Dijk 1978, Anderson and Pearson 1984, Rapp et al. 2007). Comprehension reflects the degree to which appropriate, meaningful connections are established between elements of text and the reader's prior knowledge.

Referential and causal/logical relations are particularly important in establishing coherence, by enabling readers to keep track of objects, people, events, and the relational information connecting facts and events mentioned in the text. These relations that readers must infer are not necessarily obvious. They can be numerous and complex; extend over long spans of the text; involve extensive background commonsense, social, cultural, and world knowledge; and require coordination of multiple pieces of information.

Human comprehension involves a number of different cognitive processes. Davis (1944), for instance, describes a still-relevant taxonomy of different skills tested in reading comprehension tests, and shows empirical evidence regarding performance variance across these nine different skills: knowledge of word meanings; ability to select the appropriate meaning for a word or phrase in light of its particular contextual setting; ability to follow the organization of a passage and to identify antecedents and references in it; selecting the main thought of a passage; answering questions that are specifically answered in a passage; answering questions that are answered in a passage but not in the words in which the question is asked; drawing inferences from a passage about its content; recognition of literary devices used in a passage and determination of its tone and mood; inferring a writer's purpose, intent, and point of view.

Subsequent research into comprehension examining long-term performance data of humans shows that comprehension is not a single gradable dimension, but comprises many distinct skills (for example, Keenan, Betjemann, and Olson [2008]). Most extant work examines small components of comprehension, rather than the capacity of machines to comprehend a complete discourse in its entirety.

Existing Approaches for Measuring Machine Comprehension

How can we test progress in this area? In this section, we summarize current approaches to measuring machine comprehension.

AI has a wide variety of evaluations in the form of shared evaluations and competitions, many of which bear on the question of machine comprehension. For example, TREC-8 (Voorhees 1999) introduced the question-answering track in which the participants were given a collection of documents and asked to answer factoid questions such as "How many calories are in a Big Mac?" or "Where is the Taj Mahal?" This led to a body of research in applying diverse techniques in information retrieval and structured databases to question answering and comprehension tasks (Hirschman and Gaizauskas 2001). The Recognizing Textual Entailment (RTE) Challenge (Dagan, Glickmann, and Magnini 2006) is another competition with relevance to comprehension. Given two text fragments, the task requires recognizing whether the meaning of one text is entailed by (can be inferred from) the other text. From 2004 to 2013, eight RTE Challenges were organized with the aim of providing researchers with concrete data sets on which to evaluate and compare their approaches. Neither the TREC nor the RTE competitions, however, addresses the

breadth and depth of human comprehension we seek.

One approach to testing broader-coverage machine comprehension seeks to leverage the existing diverse battery of human comprehension tests, such as SATs, domain-specific science tests, and so on (for example, Barker et al. [2004] and Clark and Etzioni [2016]). The validity of standardized tests lies in their ability to identify humans who are more likely to succeed at a certain task, such as in the practice of medicine or law.

As such, human tests are intended to effectively discriminate among intelligent human applicants, but as E. Davis (2016) notes, they do not necessarily contain classes of questions relevant to discriminating between human and artificial intelligence: questions that are easy for humans but difficult for machines, that are subjective, and so on.

Recent work on commonsense reasoning points to one possible alternative approach. The Winograd Schema Challenge (Levesque, Davis, and Morgenstern 2012; Morgenstern et al. 2016), for instance, can be seen as comprehension in a microcosm: a single story in a single sentence or very short passage with a single binary question that can in principle be reliably answered only by a system that has some commonsense knowledge. In each question there is a special word, such as that underlined in the following example, that can be replaced by an alternative word in a way that fundamentally changes the sentence's meaning.

> The trophy would not fit into the brown suitcase because it was too big/small.
> What was too big/small?
> Answer 0: the trophy
> Answer 1: *the suitcase*

In each example, the reader's challenge is to disambiguate the passage. By design, clever tricks involving word order or other features of words or groups of words will not work. In the example above, contexts where "big" can appear are statistically quite similar to those where "small" can appear, and yet the answer must change. The claim is that doing better than guessing requires readers to figure out what is going on; for example, a failure to fit is caused by one of the objects being too big and the other being too small, and readers must determine which is which.

SQUABU, for "science questions appraising basic understanding" (Davis 2016), generalizes this approach into a test-construction methodology and presents a series of test material for machines at fourth-grade and high school levels. Unlike the human counterparts of such tests, which focus on academic material, these tests focus on commonsense knowledge such as the understanding of time, causality, impossible or pointless scenarios, the human body, combining facts, making simple inductive arguments of indeterminate length, relating formal science to the real world, and so forth. Here are two example questions from SQUABU for fourth-grade level:

> Sally's favorite cow died yesterday. The cow will probably be alive again (A) tomorrow; (B) within a week; (C) within a year; (D) within a few years; (E) The cow will never be alive again.
> Is it possible to fold a watermelon?

Winograd schemas and SQUABU demonstrate some areas where standardized tests lack coverage for testing machines. Both tests, however, are entirely generated by experts and are difficult to scale to large numbers of questions and domains; neither is directed at broad-coverage comprehension.

Desiderata for a Comprehension Challenge

In a full-coverage test of comprehension, one might want to be able to ask a much broader range of questions. Suppose, for example, that a candidate software program is confronted with a just-published spy thriller, for which there are no web-searchable CliffsNotes yet written. An adequate system (Marcus 2014, Schank 2013) should be able to answer questions such as the following: Who did what to whom? Who was the protagonist? Was the CIA director good or evil? Which character leaked the secrets? What were those secrets? What did the enemy plan to do with those secrets? Where did the protagonist live? Why did the protagonist fly to Moscow? How does the story make the reader/writer feel? And so forth. A good comprehension challenge should evaluate the full breadth and depth of human comprehension, not just knowledge of common sense. To our knowledge, no previous test or challenge has tried to do this in a general way.

Another concern with existing test-construction methodology for putative comprehension challenges is the lack of transparency in the test creation and curation process. Namely, why does a test favor some questions and certain formulations over others? There is a central, often-unspoken role of the test curator in choosing the questions to ask, which is a key aspect of the comprehension task.

Given a news article, story, movie, podcast, novel, radio program, or photo — referred to as a document from this point forward — an adequate test should draw from a full breadth of all document-relevant questions with document-supported answers that humans can infer.

We suggest that the coverage goal of the comprehension challenge can be phrased as an empirical statement:

> A comprehension test should cover the full range of questions and answers that humans would expect other humans to reasonably learn or infer from a given document.

How can we move toward this goal?

The C[3] Test

We suggest that the answer begins with crowdsourcing. Previous work has shown that crowdsourcing can be instrumental in creating large-scale shared data sets for evaluation and benchmarking.

The major benefits of crowdsourcing are enabling scaling to broader coverage (for example, of domains, languages), building significantly larger data sets, and capturing broader sets of answers (Arroyo and Welty 2014), as well as gathering empirical data regarding reliability and validity of the test (Paritosh 2012).

Imagenet (Deng et al. 2009), for example, is a large-scale crowdsourced image database consisting of 14 million images with over a million human annotations, organized by the Wordnet lexicon; it has been a catalyst for recent computer vision research with deep convolutional networks (Krizhevsky, Sutskever, and Hinton 2012). Freebase (Bollacker et al. 2008) is a large database of human-curated structured knowledge that has similarly sparked research fact extraction (Mintz et al. 2009; Riedel, Yao, and McCallum 2010).

Christoforaki and Ipeirotis (2014) present a methodology for crowdsourcing the construction of tests using the questions and answers on the community question-answering site Stack Overflow.[2] This work shows that open-ended question and answer content can be turned into multiple-choice questions using crowdsourcing. Using item response theory on crowdsourced performance on the test items, they were able to identify the relative difficulty of each question.

MCTEST (Richardson, Burges, and Renshaw 2013) is a crowdsourced comprehension test corpus that consists of approximately 600 fictional stories written by Amazon Mechanical Turk crowd workers. Additionally, the crowd workers generated multiple-choice questions and their correct answers, as well as plausible but incorrect answers. The workers were given guidelines regarding the story, questions, and answers, such as that they should ask questions that make use of information in multiple sentences. The final test corpus was produced by manual curation of the resulting stories, questions, and answers. This approach is promising, as it shows that it is possible to generate comprehension tests using crowdsourcing. However, much like the standardized and commonsense tests, the test-curation process here is not entirely transparent nor generalizable to other types of documents and questions.

The question at hand is whether we can design reliable processes for crowdsourcing the construction of comprehension tests that provide us with measurable signals and guarantees of quality, relevance, and coverage, not just whether we can design a test.

As an alternative, and as a starting point for further discussion, we propose here a crowdsourced comprehension challenge (C[3]). At the root is a document-focused imitation game, which we call the *iterative crowdsourcing comprehension game* (ICCG), the goal of which is to generate a systematic and comprehensive set of questions and validated answers relevant to any given document (video, text story, podcast, or other). Participants are incentivized to explore questions and answers exhaustively, until the game terminates with an extensive set of questions and answers. The C[3] is then produced by aggregating and curating questions and answers generated from multiple iterations of the ICCG.

The structure, which necessarily depends on cooperative yet independent judgments from multiple humans, is inspired partly by Luis von Ahn's work. For example, in the two-player ESP game (von Ahn and Dabbish 2004) for image labeling, the goal is to guess what label your partner would give to the image. Once both players have typed the exact same string, they win the round, and a new image appears. This game and others in the games with a purpose series (von Ahn 2006) introduced the methodology of *input agreement* (Law and von Ahn 2009), where the goal of the participants is to try to agree on the input, encouraging them to model the other participant. The ICCG extends this to a three-person imitation game, itself partially in the spirit of Turing's original test (Turing 1950).

The Iterative Crowdsourcing Comprehension Game

The iterative crowdsourcing comprehension game (ICCG) is a three-person game. Participants are randomly assigned to fill one of three roles in each run of the game: reader (R), guesser (G), or judge (J). Players are sampled from a norming population of interest (for example, one might make tests at the second-grade level or college level). They should not know each other and should be identified only by anonymized screen names that are randomly assigned afresh in each round. They cannot communicate with each other besides the allowed game interactions.

Only the judge and the reader have access to the document (as defined earlier, text, image, video, podcast, and others); the guesser is never allowed to see it. The only thing readers and judges have in common is this document that they can both comprehend.

The purpose of the game is to generate a comprehensive set of document-relevant questions (with corresponding document-supported answers) as an outcome. The judge's goal is to identify who is the genuine document holder. The reader's goal is to prove possession of the document, by asking document-relevant questions and by providing document-supported answers. The guesser's goal is to establish possession of the document, by learning from prior questions and answers.

A game consists of a sequence of rounds, as depicted in figure 1. A shared whiteboard is used for keeping track of questions and answers, which are pub-

lished at the end of each round. The whiteboard is visible to all participants and allows the guesser to learn about the content of the document as the game proceeds. (Part of the fun for the guesser lies in making leaps from the whiteboard in order to make educated guesses about new questions.)

Each round begins with randomly assigning either the reader or the guesser to play the questioner for the round. The questioner writes down a question for this round. The reader's goal, while playing questioner, is to ask novel questions that have reliable document-supported answers. As the game proceeds, the reader is incentivized to exhaust the space of document-supported questions to be distinguished from the guesser. The reader, as questioner, does not earn points for asking questions that the guesser could answer correctly using nondocument knowledge or conclude from prior questions and answers on the whiteboard. When the questioner is the guesser, their goal is to ask revealing questions to learn as much about the story as quickly as possible.

At this point we have a question, from either the reader or guesser. The question is shared with the other participant,[3] who independently answers.

The judge is presented with both the question and the two answers with authors anonymized and attempts to identify which one is the reader. This anonymization is done afresh for the next round. The objective of both the reader and guesser is to be chosen as the reader by the judge, so both are incentivized to ask questions and generate answers that will convince the judge that they are in possession of the document.

The round is scored using this simple rubric: The judge earns a point for identifying the reader correctly, and the reader or guesser earns a point for being identified as the document holder by the judge.

At the end of each round, the question and the reader's and guesser's answers are published on the whiteboard. The reader's job is exhaustively to ask document-relevant questions, without generating questions that the guesser could extract from the accumulated whiteboard notes; the guesser's job is to glean as much information as possible to improve at guessing.

Initially, it is very easy for the judge to identify the reader. However, roughly every other round the guesser (when chosen to be the questioner) gets to ask a question and learn the reader's and judge's answers to that question. The main strategic goal of the guesser is to erode their disadvantage, the lack of access to the document, as quickly as possible. For example, the guesser might begin by asking basic information-gathering questions: *who, what, where, when, why,* and *how* questions.[4] The increased knowledge of the document revealed through the questions and answers should improve guessing performance over rounds.

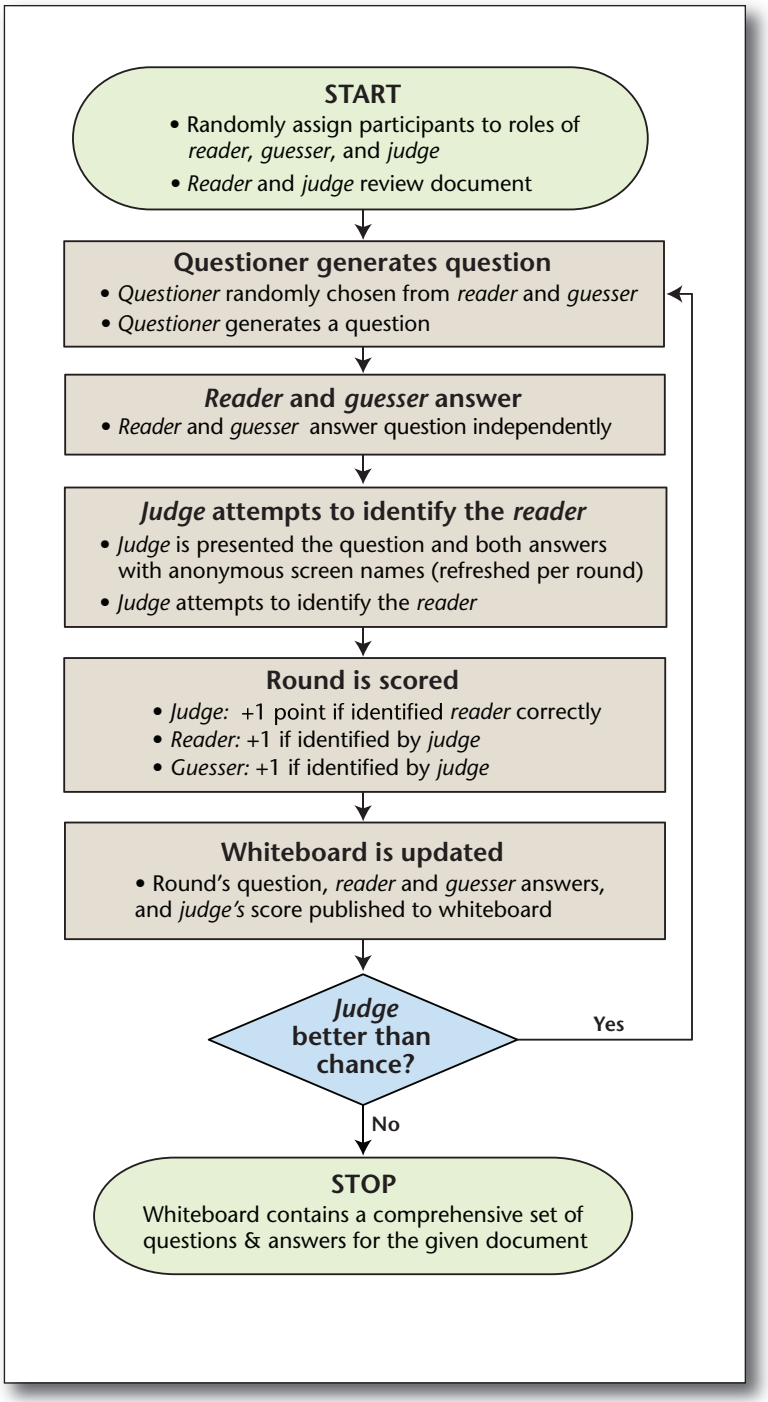

Figure 1. The Iterative Crowdsourcing Comprehension Game

The game concludes when all attempts at adding further questions fail to discriminate between the guesser and reader. This implies that the corpus of questions and answers collected on the whiteboard is a comprehensive set, that is, sufficient to provide an understanding comparable to having read the document. There can be many different sets of questions,

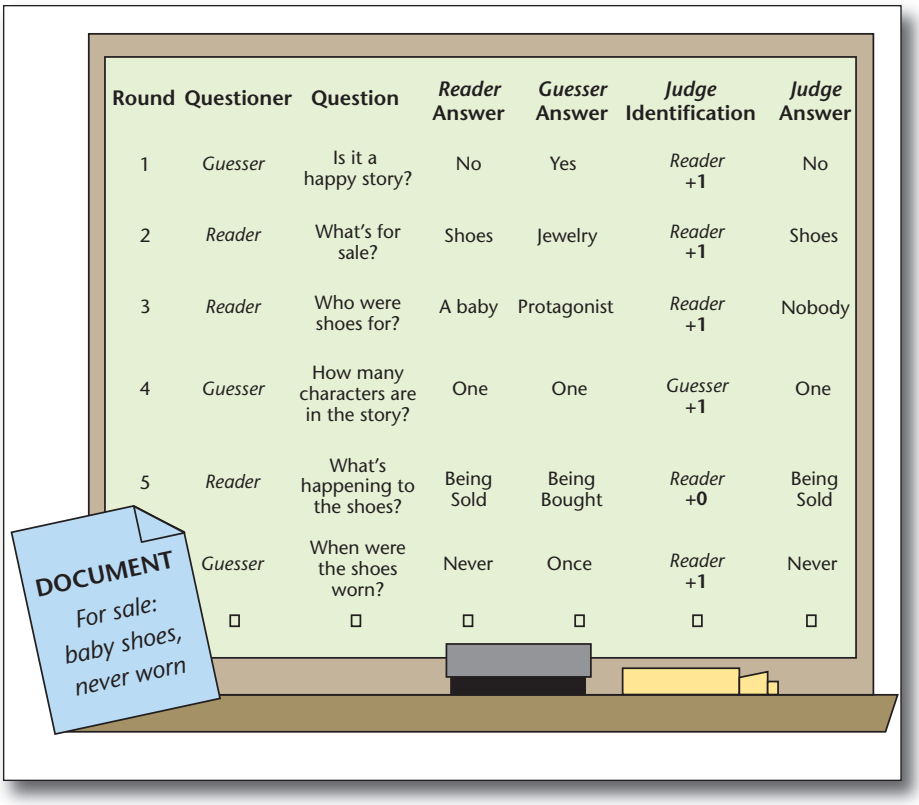

Figure 2. An Example Whiteboard.

Created for the document "For sale: baby shoes, never worn."

due to sequence effects and variance in participants. We repeat the ICCG manifold to collect the raw material for the construction of the crowdsourced comprehension challenge.

Figure 2 depicts an example whiteboard after several rounds of questioning for a simple document, a six-word novel attributed to Ernest Hemingway.

Constructing the Crowdsourced Comprehension Challenge

Given a document, each run of the game above produces a set of document-relevant questions and document-validated answers, ultimately producing a comprehensive (or at least extensive) set of questions. By aggregating across multiple iterations of the game with the same document, we obtain a large corpus of document-relevant questions and validated answers. This is the raw data for constructing the comprehension test. Finalizing the test requires further aggregation, deduplication, and filtering using crowdsourced methods, for example, the Find/Fix/Verify methodology (Bernstein et al. 2010).

This approach suggests that comprehension must be considered relative to a population. This turns our original goal for the challenge — full range of questions and answers that humans would expect other humans to reasonably learn or infer from a given document — into an empirical and crowdsourceable goal. Additionally, this allows us to design testing instruments tailored across skill levels, ages, or domains, as well as adaptable to a wide swath of cultural contexts, by sampling participants from different populations.

Figure 3 depicts the process of constructing the final test, which features the crowdsourced collection of the question–answer pairs.

Using the C^3, a broad-coverage comprehension challenge can be constructed using crowdsourcing. By varying the population, we can construct comprehension tests that reveal the comprehension of second graders or doctors. In addition, by varying the format of questions and answers, open-ended, multiple choice, Boolean, and others, or restricting allowable questions to be of a certain type, we can construct different challenges.

Conclusions and Future Work

Improved machine comprehension would be a vital step toward more general artificial intelligence and could potentially have enormous benefits for humanity, if machines could integrate medical, scientific, and technological information in ways that were humanlike.

Here we propose C^3, the crowdsourced comprehension challenge, and one candidate technique for generating such tests, the ICCG, which yield a comprehensive, relevant, and human-validated corpus of questions and answers for arbitrary content, fiction or nonfiction, presented in a variety of forms. The game also produces human-level performance data for constructing tests, which with suitable participants (such as second graders or adult native speakers of a certain language) could be used to yield a range of increasingly challenging benchmarks. It could also be tailored to specific areas of knowledge and inference (for example, the domain of questions could be restricted to commonsense understanding, to science or medicine, or to cultural and social understanding). Unlike specific tests of expertise, this is a general test-generation procedure whose scope is all questions that can be reliably answered by humans (either in general, or drawn from a population of interest) holding the document.

Of course, more empirical and theoretical work is needed to implement, validate, and refine the ideas proposed here. Variations of the ICCG might be useful for different data-collection processes (for example, Paritosh [2015] explores a version where the individual reader and guesser are replaced by samples of readers and guessers). An important area of future work is the

design of incentives to make the game more engaging and useful (for example, Prelec [2004]). We believe that crowdsourced processes for the design of human-level comprehension tests will be an invaluable addition to the arsenal of assessments of machine intelligence and will spur research in deep understanding of language.

Acknowledgments

The authors would like to thank Ernie Davis, Stuart Shieber, Peter Norvig, Ken Forbus, Doug Lenat, Nancy Chang, Eric Altendorf, David Huynh, David Martin, Nick Hay, Matt Klenk, Jutta Degener, Kurt Bollacker, and participants and reviewers of the Beyond the Turing Test Workshop at AAAI 2015 for insightful comments and suggestions on the ideas presented here.

Notes

1. This is part of the VisualGenome corpus, visualgenome.org.

2. stackoverflow.com.

3. One might also secure an answer from the judge, as a validity check and to gain a broader range of acceptable answers (for example, *shoes* or *baby shoes* might both work for a question about the Hemingway story shown in figure 2).

4. The popular Twenty Questions game is a much simpler version, where the guesser tries to identify an object within twenty yes/no questions. Questions such as "Is it bigger than a breadbox?" or "Does it involve technology for communications, entertainment, or work?" allow the questioner to cover a broad range of areas using a single question.

References

Anderson, R. C., and Pearson, P. D. 1984. A Schema-Theoretic View of Basic Processes in Reading Comprehension. *Handbook of Reading Research* Volume 1, 255–291. London Routledge.

Barker, K.; Chaudhri, V. K.; Chaw, S. Y.; Clark, P.; Fan, J.; Israel, D.; Mishra, S.; Porter, B.; Romero, P.; Tecuci, D,; and Yeh, P. 2004. A Question-Answering System for AP Chemistry: Assessing KR&R Technologies, 488–497. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Ninth International Conference*. Menlo Park, Calif: AAAI Press.

Bernstein, M. S.; Little, G.; Miller, R. C.; Hartmann, B.; Ackerman, M. S.; Karger, D. R.; Crowell, D.; and Panovich, K. 2010. Soylent: A Word Processor with a Crowd Inside. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology,* 313–322. New York: Association for Computing Machinery. dx.doi.org/10.1145/1866029.1866078

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge, 1247–1250. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008*. New York: Association for Computing Machinery. dx.doi.org/10.1145/1376616.1376746

Christoforaki, M., and Ipeirotis, P. 2014. STEP: A Scalable Testing and Evaluation Platform. In *Proceedings of the Second AAAI Conference on Human Computation and Crowdsourcing*. Palo Alto, CA: AAAI Press.

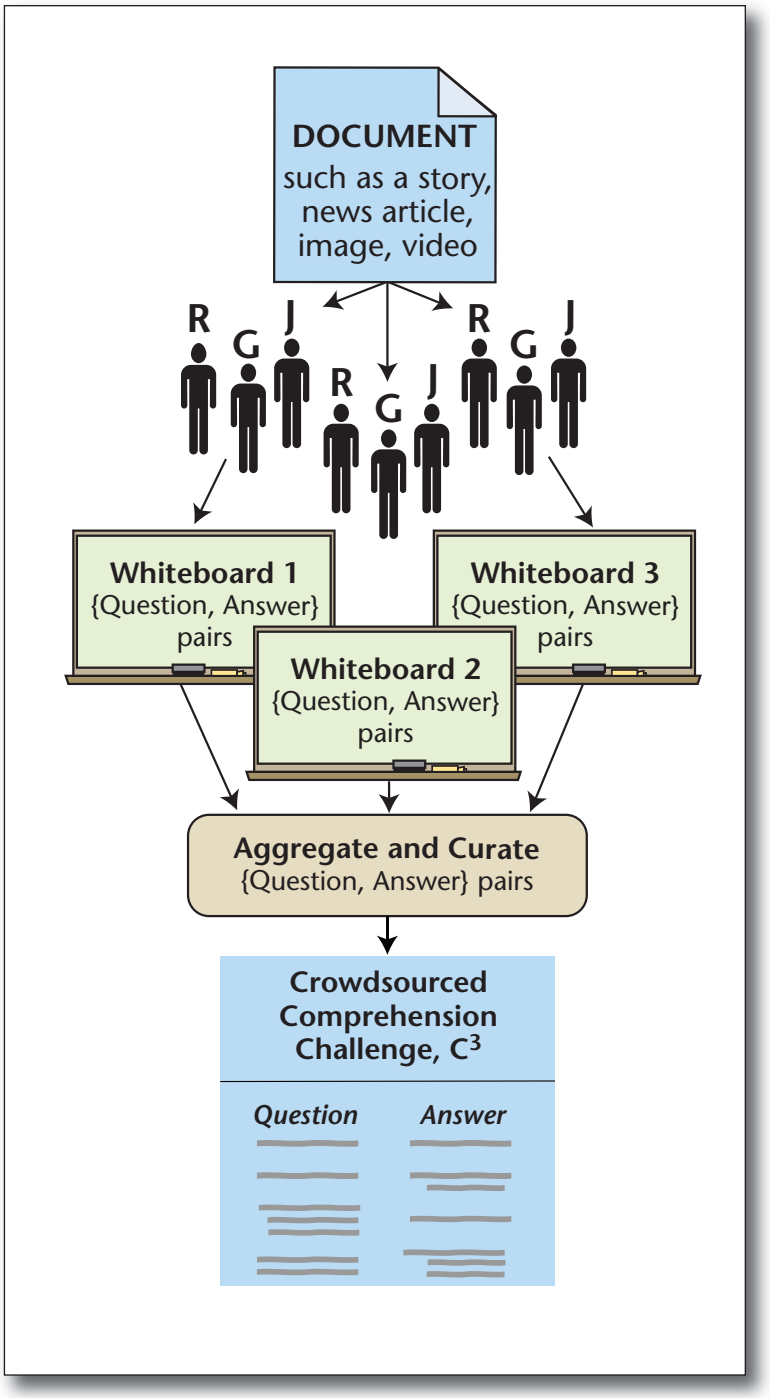

Figure 3. Crowdsourced Comprehension Challenge Generation.

Clark, P., and Etzioni, O. 2016. My Computer Is an Honor Student — But How Intelligent Is It? Standardized Tests as a Measure of AI. *AI Magazine* 37(1).

Dagan, I.; Glickman, O.; and Magnini, B. 2006. The PASCAL Recognising Textual Entailment Challenge. In *Machine*

Learning Challenges, Lecture Notes in Computer Science Volume 3944, 177–190. Berlin: Springer. dx.doi.org/10.1007/11736790_9

Davis, E. 2016. How to Write Science Questions That Are Easy for People and Hard for Computers. *AI Magazine* 37(1).

Davis, F. B. 1944. Fundamental Factors of Comprehension in Reading. *Psychometrika* 9(3): 185–197. dx.doi.org/ 10.1007/BF02288722

Deng, J.; Dong, W.; Socher, R.; Li, L. J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A Large-Scale Hierarchical Image Database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 248–255. Piscataway, NJ: Institute of Electrical and Electronics Engineers. dx.doi.org/10.1109/CVPR.2009. 5206848

Hirschman, L., and Gaizauskas, R. 2001. Natural Language Question Answering: The View from Here. *Natural Language Engineering* 7(04): 275–300. dx.doi.org/10.1017/S1351324901002807

Keenan, J. M.; Betjemann, R. S.; and Olson, R. K. 2008. Reading Comprehension Tests Vary in the Skills They Assess: Differential Dependence on Decoding and Oral Comprehension. *Scientific Studies of Reading* 12(3): 281–300. dx.doi.org/10.1080/10888430802132279

Kintsch, W., and van Dijk, T. A. 1978. Toward a Model of Text Comprehension and Production. *Psychological Review* 85(5): 363. dx.doi.org/10.1037/0033-295X.85.5.363

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012,* 1097–1105. La Jolla, CA: Neural Information Processing Systems Foundation, Inc.

Law, E., and von Ahn, L. 2009. Input-Agreement: A New Mechanism for Collecting Data Using Human Computation Games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 1197–1206. New York: Association for Computing Machinery. dx.doi.org/10.1145/1518701.1518881

Levesque, H.; Davis, E.; and Morgenstern, L. 2012 The Winograd Schema Challenge. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference* (KR2012), 552–561. Palo Alto, CA: AAAI Press.

Marcus, G. 2014. What Comes After the Turing Test? *New Yorker* (June 9).

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 2009. Distant Supervision for Relation Extraction Without Labeled Data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics,* 1003–1011. Stroudsburg, PA: Association for Computational Linguistics. dx.doi.org/10.3115/1690219.1690287

Morgenstern, L.; Davis, E.; Ortiz, C. L. Jr. 2016. Planning, Executing, and Evaluating the Winograd Schema Challenge. *AI Magazine* 37(1).

Paritosh, P. 2012. Human Computation Must Be Reproducible. In *CrowdSearch 2012: Proceedings of the First International Workshop on Crowdsourcing Web Search*. Ceur Workshop Proceedings Volume 842. Aachen, Germany: RWTH-Aachen University.

Paritosh, P. 2015. Comprehensive Comprehension: A Document-Focused, Human-Level Test of Comprehension. Paper presented at Beyond the Turing Test, AAAI Workshop W06, Austin TX, January 25.

Poggio, T., and Meyers, E. 2016. Turing++ Questions: A Test for the Science of (Human) Intelligence. *AI Magazine* 37(1).

Prelec, D. 2004. A Bayesian Truth Serum for Subjective Data. *Science* 306(5695): 462–466. dx.doi.org/10.1126/science.1102081

Rapp, D. N.; Broek, P. V. D.; McMaster, K. L.; Kendeou, P.; and Espin, C. A. 2007. Higher-Order Comprehension Processes in Struggling Readers: A Perspective for Research and Intervention. *Scientific Studies of Reading* 11(4): 289–312. dx.doi.org/10.1080/10888430701530417

Richardson, M.; Burges, C. J.; and Renshaw, E. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *EMNLP 2013: Proceedings of the Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics.

Riedel, S.; Yao, L.; and McCallum, A. 2010. Modeling Relations and Their Mentions Without Labeled Text. In *Machine Learning and Knowledge Discovery in Databases: Proceedings of the European Conference, ECML PKDD 2010*. Lecture Notes in Artificial Intelligence Volume 6322, 148–163. Berlin: Springer. dx.doi.org/10.1007/978-3-642-15939-8_10

Saygin, A. P.; Cicekli, I.; and Akman, V. 2003. Turing Test: 50 Years Later. In *The Turing Test: The Elusive Standard of Artificial Intelligence,* ed. J. H. Moor, 23–78. Berlin: Springer. dx.doi.org/10.1007/978-94-010-0105-2_2

Schank, R. P. 2013. *Explanation Patterns: Understanding Mechanically and Creatively.* London: Psychology Press.

Shieber, S. M. 1994. Lessons from a Restricted Turing Test. *Communications of the ACM* 37(6): 70–78. dx.doi.org/10.1145/175208.175217

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind* 59(236): 433–460. dx.doi.org/10.1093/mind/LIX.236.433

von Ahn, L. 2006. Games with a Purpose. *Computer* 39(6): 92–94. dx.doi.org/10.1109/MC.2006.196

von Ahn, L., and Dabbish, L. 2004. Labeling Images with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 319–326. New York: Association for Computing Machinery. dx.doi.org/10.1145/985692.985733

Voorhees, E. M. 1999. The TREC-8 Question Answering Track Report. In *Proceedings of The Eighth Text Retrieval Conference, TREC 1999,* 77–82. Gaithersburg, MD: National Institue of Standards and Technology.

Poggio, T., and Meyers, E. 2016. Turing++ Questions: A Test for the Science of (Human) Intelligence. *AI Magazine* 37(1).

Zitnick, C. L.; Agrawal, A.; Antol, S.; Mitchell, M.; Batra, D.; Parikh, D. 2016. Measuring Machine Intelligence Through Visual Question Answering. *AI Magazine* 37(1).

Praveen Paritosh is a a senior research scientist at Google leading research in the areas of human and machine intelligence. He designed the large-scale human-machine curation systems for Freebase and the Google Knowledge Graph. He was the co-organizer and chair for the SIGIR WebQA 2015 workshop, the Crowdsourcing at Scale 2013, the shared task challenge at HCOMP 2013, and Connecting Online Learning and Work at HCOMP 2014, CSCW 2015, and CHI 2016 toward the goal of galvanizing research at the intersection of crowdsourcing, natural language understanding, knowledge representation, and rigorous evaluations for artificial intelligence.

Gary Marcus is a professor of psychology and neural science at New York University and chief executive officer and founder of Geometric Intelligence, Inc. He is the author of four books, including *Kluge: The Haphazard Evolution of the Human Mind* and *Guitar Zero*, and numerous academic articles in journals such as *Science* and *Nature*. He writes frequently for *The New York Times* and *The New Yorker,* and is coeditor of the recent book, *The Future of the Brain: Essays By The World's Leading Neuroscientists*.