Considerations for Evaluating Models of Language Understanding and Reasoning

Gabriel Recchia
University of Cambridge
Cambridge, UK CB3 9DT
glr29@cam.ac.uk

Abstract

Efforts to construct tasks for evaluating reasoning systems face a tradeoff between ecological validity and interpretability. That is, as task difficulty and diversity increases, the easier it is to design a set of tasks whose solution requires a general-purpose algorithm capable of generalizing to many different scenarios, but the harder it becomes for the researcher to understand what information an algorithm needs in order to perform well on the task. The Facebook bAbI dataset [1] represents one recent attempt to design such an evaluation framework that preserves experimental control while allowing the difficulty of the task to be incrementally scaled up. Although this is a very promising approach, some tasks in this dataset may not fully capture the cognitive skills intended, while others may not be soluble by even sophisticated human reasoners without prior knowledge (unavailable in the training data). A complementary task framework and evaluation dataset modeled closely on [1] is presented, which arguably preserves experimental control and allows for difficulty to be scaled up incrementally while also ensuring that all information relevant to solving the tasks is preserved in the training data.

1 Introduction

Although the use of "artificial worlds" to evaluate reasoning and natural language understanding systems has historically been associated with the brittle, difficult-to-scale approaches pursued in the 1970s (e.g. [2,3]), there are many examples within the machine learning literature of artificial subproblems that clarified key challenges to be overcome and led to major advances [4,5,6]. In particular, early advances in neural nets owe much to synthetic data, such as many of the inputs used in Rumelhart & McClelland's parallel distributed processing framework [7] and the toy language employed in Elman's early work on simple recurrent nets [5]. Bordes et al. [8] have pointed out that in contrast to the toy worlds used by early scripted AI systems based on symbolic logic, most of the success stories for synthetic data have involved cases in which a *learning algorithm* was applied to synthetic data. As such, they advocate the use of synthetic datasets in the context of the traditional machine learning paradigm: a task consists of a training set and a test set, each of which is generated by the same algorithm. If a sufficiently diverse set of such tasks is provided, a reasoning system that performs well on all of them may be likely to perform well on other tasks involving reasoning and natural language understanding. In addition, the experimenter retains full control over and knowledge of the process that generated the training and test data. The Facebook bAbI dataset [1,8] is a novel battery of twenty tasks consisting of synthetic training and testing data, with goals of being (1) *self-contained*: that is, all of the information necessary to perform well at the task should be present within the training data; and (2) *wide-coverage*: the tasks in the dataset are intended to correspond to diverse cognitive abilities (path-finding, positional reasoning, use of negation, coreference resolution, etc.). It is this combination of goals that distinguishes [1] from most other efforts

to build question-answering datasets for evaluating systems for text comprehension or general reasoning. For example, ARISTO [9], MCTest [10], the Winograd Schema Challenge [11] and many "visual" QA systems (e.g. [12,13]) are wide-coverage but not self-contained. Conversely, efforts such as the Synthetic Visual Reasoning Test [14], Bongard problems [15], and Raven's Progressive Matrices are self-contained but not wide-coverage, as they focus on visual reasoning only and do not involve natural language comprehension.

Training data in the bAbI dataset is presented in the form of 1,000 items (e.g. Fig. 1) per task. The dataset also includes a "shuffled" variant of each task in which each character is replaced with another (e.g., all 'h's replaced with 'z's). For researchers who evaluate systems on the dataset as intended— using only the training data and no knowledge derived from external sources—this shuffling should make little difference. The ability to perform well on all tasks is hypothesized in [1] to be necessary but not sufficient for any system capable of human-level natural language understanding.

1 Julie went to the park this morning.
2 Julie moved to the kitchen yesterday.
3 Where was Julie before the park? kitchen 1 2
4 Julie moved to the bedroom this afternoon.
5 Mary journeyed to the school yesterday.
6 Where was Julie before the bedroom? park 4 1

Figure 1: Example item from bAbI task 14 (time reasoning).

2 Advantages and Limitations

The bAbI dataset is unique in its goals of constructing a dataset for evaluating general-purpose reasoning systems that is simultaneously self-contained and wide-coverage. We find this compelling because it permits the kinds of tasks presented to any given model to be to be incrementally scaled up in difficulty and complexity, which could help to provide the research community with a metaphorical 'smooth gradient' on which to develop algorithms capable of achieving human performance on increasingly difficult and diverse tasks. Each bAbI task is described with a title such as 'compound coreference' or 'time manipulation,' reflecting the creators' goal to "categorize different kinds of questions into skill sets, which become our tasks" [1]. However, the initial bAbI tasks do not always fully capture the cognitive skills described by their titles. For example, task 18 can readily be solved with a symbolic matching process that bears little relation to what a cognitive psychologist would describe as "size reasoning."[1] Furthermore, some tasks may be much more difficult than intended. The "shuffled" variant described earlier clearly demonstrates the difficulty of the bAbI tasks for any algorithm, but also illustrates the difficulty of the task for humans. If we consider the shuffled, weakly supervised "three supporting facts" task, for example, it is difficult to imagine how a person could obtain excellent performance without knowledge that some of the symbols under investigation refer to objects with persistent states, that other symbols indicate actions which transform the states of these objects, and with no feedback about the possible states or transformations. Given the amount of research demonstrating that joint attention to language and perceived objects/events/properties is critical to children's language learning [16] and that computational models of human language learning benefit from joint presentation of language and perceptual information [17,18], providing training data without a sensory representation of the state of the world (or at least a symbolic representation thereof) may be making the algorithm's task even more difficult than language learning is for humans.

We believe researchers would have more confidence that the weakly supervised bAbI tasks could be solved with a general algorithm if it were more obvious that a person with no understanding of the language involved could successfully solve each task. To this end, we release GABITS (the Grounded and bAbI-Inspired Task Set), a set of several "weakly

[1] Specifically, consider replacing the phrase "fit in" or "fits inside" in bAbI task 18 with the word "outclasses", and the phrase "bigger than" with the phrase "outclassed by." Nearly any pattern-matching algorithm capable of solving the original task would be able to solve this slightly modified task that makes no reference to size. The term "size reasoning," therefore, does not quite seem to capture the cognitive skills that are actually required to solve the task, which seem more related to the ability to handle transitivity.

supervised" tasks similar to those in the Facebook bAbI dataset, along with very simple perceptual and symbolic representations of the stories being described. In this way, we hope to have released a dataset that is difficult enough that its solution requires a fairly general reasoning algorithm, yet which is also more obviously self-contained. The tasks and associated documentation may be found at http://nowin2d.com/gabits/, and are described briefly in the following section.

3 The GABITS dataset

GABITS is a set of tasks for evaluating natural language understanding and reasoning systems. It is very much inspired by [1], and strives to present a complementary set of tasks grounded in synthetically generated images. A symbolic representation of the environment is also provided for researchers who are uninterested in building models of perception, but who are interested in building models that jointly attend to the state of the environment and natural language. A training instance consists of an unlabeled image (optional but recommended), a symbolic representation (optional), and a narrative with associated questions and answers. An example unlabeled image and narrative are shown in Figure 2. Images and symbolic representations of the environment are not available at test time. As with the Facebook tasks, all training and test instances are synthetically generated, and researchers are encouraged to use the same learner across tasks and not to resort to task-specific heuristics. By incrementally increasing the diversity of the questions in future versions, we hope that the evaluation will become incrementally more difficult and that the algorithms required to solve it will become more general in their capabilities.

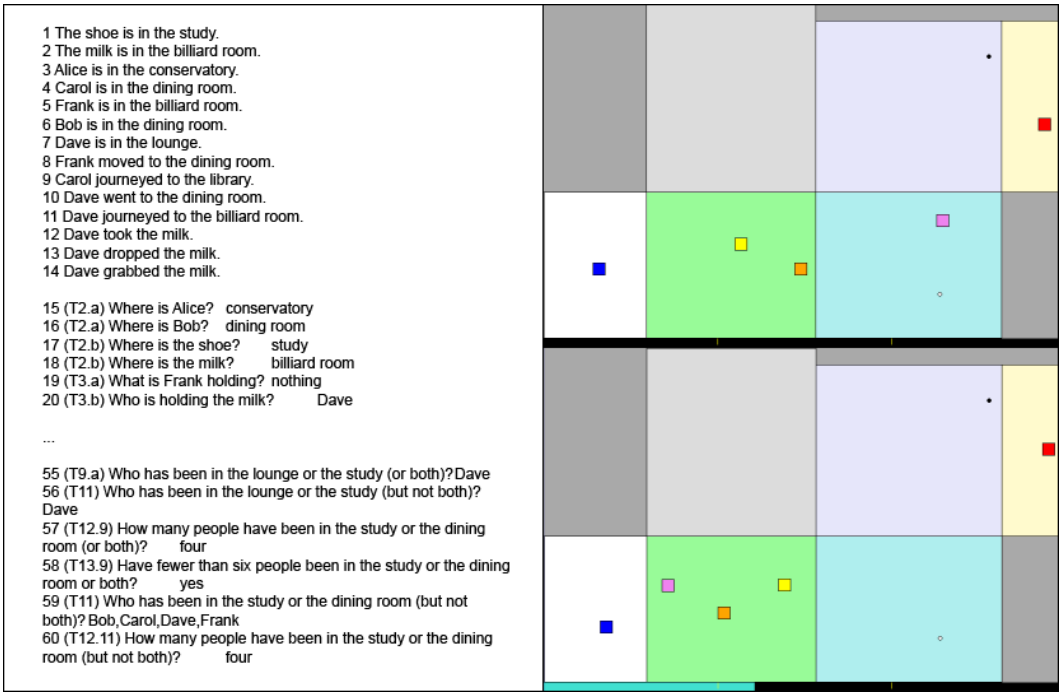

Figure 2: Partial example of a training item from GABITS. Left: an example narrative (lines 1-14) is followed by several questions that range from easy (15-20) to hard (55-60). Right: the first two frames of the images accompanying the narrative. In the second frame, the magenta box ("Frank") has moved from the blue area ("the billiard room") to the green area ("the dining room"), reflected in line 8.

Questions in GABITS are not explicitly identified with cognitive abilities, but are rather placed into groups and subgroups by characteristics of the question. For example, questions in group T2 ("Task 2") pertain to location, with questions in group T2.a asking about the location of a person and group T2.b asking about the location of an item. Questions

represent a range of difficulty levels. For example, questions in group T4 concern which objects individuals have held over the course of the narrative, with T4.a. asking for the set of items that have been held by a particular agent and T.4.b. asking for the set of agents that have held a particular item. Questions in group T12.4 ask how many items have been held by a particular agent or how many agents have held a particular item. Questions in T13.4 ask whether or not the number of agents who have held a particular item, or items held by some agent, is greater or lesser than some number. Despite the fact that all information required to solve the tasks is present in the training data, the wide range of difficulty levels suggests that the GABITS dataset may offer a challenge to natural language understanding systems for some time to come.

4 Conclusion

As the complexity or difficulty of an evaluation task increases, an initially self-contained task may inadvertently lose this quality. The present work has illustrated that although the bAbI tasks have many advantages, high performance does not always require the cognitive ability described in the task's title, and it is not clear that human reasoners would be capable of high performance on the (shuffled) weakly supervised version of each task. By releasing GABITS, we hope to have addressed some of the bAbI dataset's limitations while retaining many of its strengths.

References

[1] Weston, J., et al. (2015). Towards AI-complete question answering: a set of prerequisite toy tasks. *arXiv* preprint:1502.05698. Data avail. https://research.facebook.com/researchers/1543934539189348

[2] Schank, R. & Abelson, R. (1975). *Scripts, plans, goals, and understanding.* Hillsdale, NJ: Lawrence Erlbaum Associates.

[3] Winograd, T. (1972). *Understanding natural language.* New York: Academic Press.

[4] Hinton, G. E. (1986). Learning distributed representations of concepts. In Proc. *CogSci*, 1, 12.

[5] Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2), 179-211.

[6] Zhou, D., et al. (2004). Learning with local and global consistency. *Advances in neural information processing systems*, *16*(16), 321-328.

[7] Rumelhart, D. E., & McClelland, J. L. (1988). *Parallel distributed processing*. USA: IEEE.

[8] Bordes, A., et al. (2015). Artificial tasks for artificial intelligence. Keynote presentation. *ICLR*.

[9] Clark, P. (2015). Elementary school science and math tests as a driver for AI: Take the Aristo Challenge. In Proc. *IAAI-15*.

[10] Richardson, M., Burges, C., & Renshaw, E. (2013). MCTest: A challenge dataset for the open-domain machine comprehension of text. In Proc. *EMNLP*, pp. 193-203.

[11] Levesque, H. J, Davis, E., & Morgenstern, L. (2011). The Winograd schema challenge. In Proc. *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

[12] Lin, X. & Parikh, D. (2015). Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. *CVPR*.

[13] Antol, S., et al. (2015). VQA: Visual question answering. In Proc. *ICCV*.

[14] Fleuret, F., et al. (2011). Comparing machines and humans on a visual categorization test. *PNAS*, 108(43), 17621-17625.

[15] Bongard, M. M. (1970). *Pattern recognition.* Rochelle Park, N.J.: Hayden, Spartan Books.

[16] Baldwin, D. A. (2014). Understanding the link between joint attention and language. In Moore, C. & Dunham, P. (Eds.), *Joint attention: its origins and role in development* (pp. 131-158). Hove, UK: Psychology Press.

[17] Howell, S. R., Jankowicz, D., & Becker, S. (2005). A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *J. Mem. Lang.*, *53*(2), 258-276.

[18] Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psych. Review, 116*(3), 463.