Question asking as program generation

computation and cognition lab // new york university

Anselm Rothe[1], Brenden Lake[1,2], & Todd Gureckis[1]

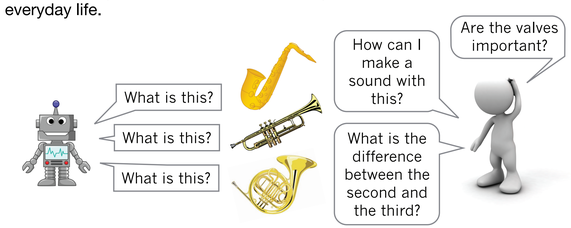
[1]Department of Psychology, [2]Center for Data Science, New York University

IF YOU COULD ASK ANYTHING, WHAT WOULD YOU ASK? 1

A question is an expression used to make a **request for information**.

Nearly all progress in Active Learning has been made with focus on a simple type of questions (label queries).

However, people use a much richer set of questions to obtain information in everyday life.

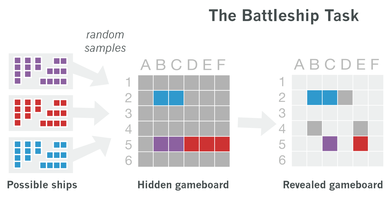


We propose a **computational framework** of how people **generate** rich questions, treating question asking as program synthesis.

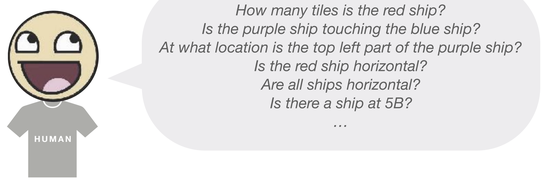
A long term goal is to develop algorithms with a human-like capacity to learn by asking rich questions. A second goal is to understand more about the computational aspects of human question asking.

QUESTION DATA SET 2

We used the human question data set from Rothe, Lake, and Gureckis (2016) with 605 questions across 18 game contexts.



Identify the hidden gameboard!

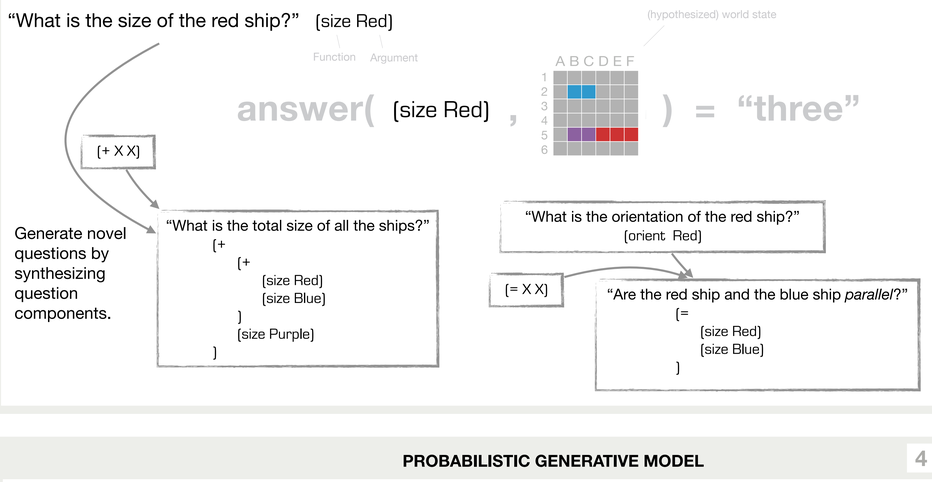


Only questions with a one-word answer were allowed and no combination of questions.

This research was supported by NSF grant BCS-1255538, the John Templeton Foundation Varieties of Understanding project, a John S. McDonnell Foundation Scholar Award to TG, and the Moore-Sloan Data Science Environment at NYU.

QUESTIONS AS PROGRAMS 3

We view questions as programs that, when executed on the state of the world output an answer.



PROBABILISTIC GENERATIVE MODEL 4

The model should be capable of asking **novel questions** in **new contexts**.

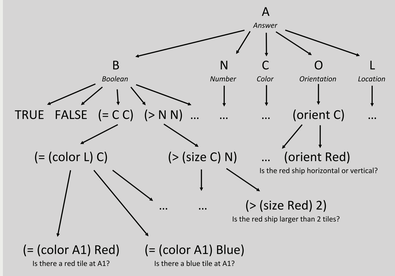
The model aims to predict **which questions people will ask**.

We used **energy-based learning** to fit the relative importance of question features (using the human question data set as training data).

GRAMMAR FOR QUESTIONS

Rules (subset)

A → B	B → TRUE	N → (size C)	C → Blue
A → N	B → FALSE	N → 1	C → Red
A → C	B → (> N N)	N → 2	C → Purple
…	…	N → 3	…
		…	



← The **space of questions** X is defined by a grammar.

Features

f_1 **Informativeness** Expected Information Gain
f_2 **Complexity** log probability under the probabilistic grammar
f_3 **Answer type** Boolean, Number, Color, Location
f_4 **Relevance** Auxiliary feature to filter out questions that do not address the game board

Optimization We had to approximate the gradient (via importance sampling) since the set of all questions X is intractably large.

Equations The energy $\mathcal{E}$ of question x is a weighted sum of its features and is related to the probability of asking x.

$$\mathcal{E}(x) = \theta_1 f_1(x) + \theta_2 f_2(x) + \ldots + \theta_K f_K(x)$$

$$p(x;\boldsymbol{\theta}) = \frac{\exp(-\mathcal{E}(x))}{\sum_{x \in X} \exp(-\mathcal{E}(x))}$$

And the objective function with question d from the human question data set.

$$\arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log p(d^{(i)};\boldsymbol{\theta})$$

RESULTS 5

QUANTITATIVE RESULTS

A We fit 4 **model variants** to all but one context and let it predict the remaining one. The log-likelihoods of the human questions were averaged across held out contexts.

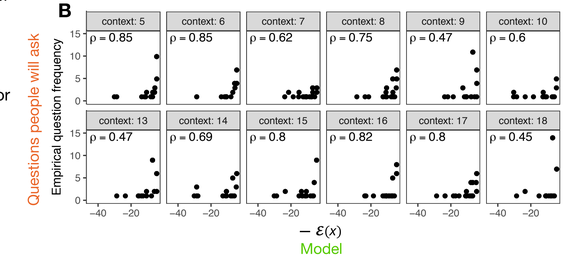
A

Model	Features	Log-likelihood
Full	**all**	**-1400.06**
Information-agnostic	not f_1	-1464.65
Complexity-agnostic	not f_2	-22993.38
Type-agnostic	not f_3	-1419.26

Models that had one key feature lesioned achieved a lower log-likelihood than the full model using all features.

B The predictions of the **full model** showed strong alignment with the question frequencies in the data set for some contexts and more modest alignment for others (average correlation ρ = .64).

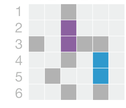
B



QUALITATIVE RESULTS

The full model was also able to generate **novel, "human-like" questions** that no human had asked.

We removed duplicate questions that were equivalent to the human questions (determined by the mutual information of their answer vectors, as well as functional equivalence up to a swapping of the arguments, e.g. *(size Blue)* is form equivalent to *(size Red)*).

Context	Question (manual translation)	Program (samples from model)
	What is the row of the bottom right red tile?	(rowL (bottomright (coloredTiles Red)))
	What is the row of the top left red tile?	(rowL (topleft (coloredTiles Red)))
	How many tiles have the same color as tile 2F?	(setSize (coloredTiles (color 2F)))
	Is tile 5F a water tile?	(isSubset (coloredTiles Water) (coloredTiles (color 5F)))
	What is the column of the top left of the tiles that have the color of the bottom right corner of the board?	(colL (topleft (coloredTiles (color (bottomright (set 1A … 6F))))))
	What is the column of the bottom right water tile?	(colL (bottomright (coloredTiles Water)))
	What is the column of the top left water tile?	(colL (topleft (coloredTiles Water)))
	How many tiles have the same color as the bottom right tile of the board?	(setSize (coloredTiles (color (bottomright (set 1A … 6F)))))
	What is the bottom right tile that has the same color as the tile 1A?	(bottomright (coloredTiles (color (topleft (set 1A … 6F)))))
	Are all the ships oriented horizontally?	(all (map (lambda x (== H (orient x))) (set Blue Red Purple)))

TAKE HOME POINTS

Our model can produce **interesting and informative questions** that people could have but did not ask in the question data set.

Our model predicts to a certain degree what questions people will ask in a given game context.

The **compositionality** of our approach is important as about 15% of the human questions did only appear in a single game context. Any model unable to synthesize novel questions would be guaranteed to fail at these 15%.