Use Mass Responses to Trivia Questions to Explore Collective Understanding of Concepts

Final Report

Yaodong Jia, Matt Fox, Justin Melnick, Imre Patyi
Georgia Institute of Technology
Atlanta, GA 30332, USA

Introduction and Motivation

"A trivia game is one where the competitors are asked questions about interesting but unimportant facts in many subjects" [1]. People all over the world and throughout the ages have played trivia games ranging from game nights in local pubs to Who Wants to Be a Millionaire on television. While mass viewership has been a frequent feature of trivia games, simultaneous mass participation is recent.

Our goal is to study patterns in the effect of question formation, grammatical complexity, difficulty of topics, etc on people's ability to provide correct answers to brief factual questions. We have a collection of data on the daily HQ Trivia games. It includes the text of the questions (12-15 per game), correct/incorrect answers, number of people choosing each of the three alternative answers, and some privately collected mark-ups on the categories of the questions posed.

Problem Definition

No data currently exists that offers insight into millions of instantaneous responses to the same question. Using statistical tools and unstructured text analytics we will parse questions and answer choices (coupled with response data) to discover patterns about how people understand certain topics as well as how to write a clear question many people will understand.

Conclusions about how people answer questions and confirm their understanding of a subject can be useful to marketing companies hoping to influence people. Moreover, social patterns may be exposed to find gaps in topical understanding among the general public. Finally, trivia fans will benefit from added insight on whether a question will be difficult before the answer is revealed.

Survey of Related Works

Some but not many works published are specific to trivia games, mostly focusing on the question aspect of the game [16]. Among the works many use text mining approaches to predict question difficulty [2], analyze semantic components of questions [3] and estimate class/category of questions [4], while some others interviewed a professional writer of questions directly [5] to understand how questions are being assembled. Though our project focuses more on the joint dynamics between question features and player responses, these works certainly provide good perspectives on how to model the data.

Given that in many trivia datasets answers are multiple choice, Fagley's observation on positional response bias [10] suggests that position and length of answers may bring bias to the accuracy-question model and need to be controlled.

As this project involves Natural Language Processing (NLP) analysis, linguistic and lexical databases such as Text8 [6], Merriam-Webster Dictionary [7] and text analysis tools such as TF-IDF [8] and Word2Vec [9, 17] will be used for decomposing question texts into features.

Another domain of research interests us is the text mining for non-adversarial question-answer structures such as in Community Question Answering services (QAS) and in online forums. Heilman [11] proposes an algorithm for generating factual questions from articles which could be useful to us. Cai, et al. [12] give a few pointers on how to identify the topic of the question from the text of the question. Gideon, et al. [13] model associations between questions and users. Eyecioglu et al. [14] and Karampatsis [15] show that support vector machine (SVM) can be profitably applied to identifying paraphrases with only light preprocessing and tagging of the data.

Data Context

The data is collected from HQ Trivia, a popular mobile trivia game. This dataset contains 2,453 questions collected from the game between 10/19/2017 and 3/28/2018, with some incorrectly entered records discarded. The game is structured as multiple rounds, in each round a question and three answer choices are presented to players and a selection must be made within 10 seconds. Only players who answer correctly may enter the next round and winners of the final round will be rewarded with cash incentives. Most times there are 12 rounds, but occasionally the game extends to 15 rounds.

We spent roughly $575 on data collection through Amazon Mechanical Turk by hiring collectors to capture data. Additionally, we host a website to showcase our analyses, which has minimal cost.

Proposed Method

In the context of this project, we consider the interaction between participants and the questions, specifically the question-choice pairs, are the proxies through which we shall study the participant-concept pairs. Because the domain for "concepts" is unbounded, we need to start with the dataset, extract features of interest, apply analysis on the features, observe analyzed results and then consolidate them to a limited number of conclusions based on the dataset. In other words, our analysis should converge to interconnected observations rather than diverge to many uncorrelated findings. For such reason, we are proposing the following analytical pipeline which reduces to few but reliable conclusions.

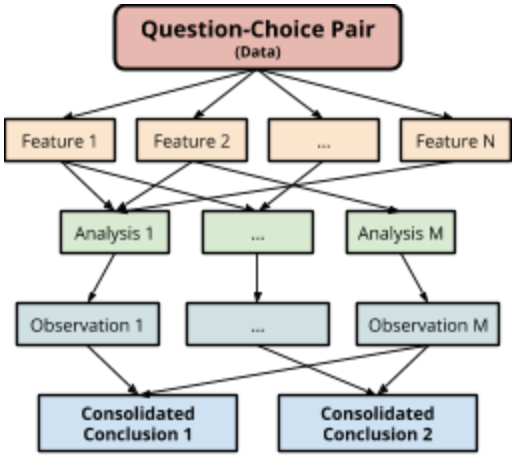


Figure 1.1.1. Proposed data analytical pipeline.

Given the data, we plan to investigate the relationships between correct_percent and various attributes, relationships between question text and others, and relationships between answer choices and others. (features)

We plan to apply traditional statistical tools, classification/regression methods and unstructured text methods to analyze the relationships and report the observations. (analysis and observations)

We will test our conclusions based on their statistical significance and if they support each other collectively. (conclusions & evaluations)

Analysis and Observations

Because the game provider needs to pay final winners of each game cash rewards while still attracting players to play, it is assumed that the

questions would have increasing difficulty each round. However, as shown in the histogram below, the percentages of correct responses are found to follow a bimodal distribution. There is one high-density distribution to the right which represents questions that answered correctly by nearly all participants, and another high-variance distribution which centers near 0.4 correct-answer-ratio.

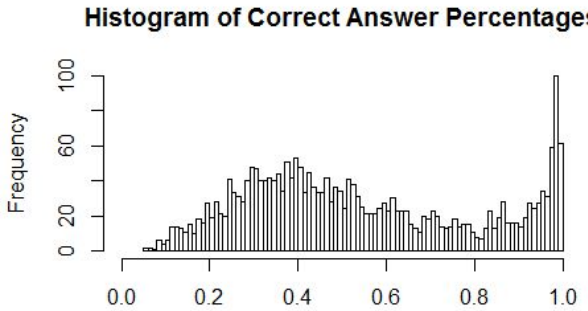


Figure 2.1.3. A bimodal distribution of correct_percent.

This finding shows that, assuming the questions represent a fair test on a random knowledge, the participants collectively share a good understanding for a range of "core" knowledge but individually display variable levels of understanding for knowledge beyond that.

To cluster the questions into "core" vs "non-core" buckets which follow different correct percentage distributions, we build an expectation-maximization algorithm to approximate the distributions:

1. Select two distributions and their parameter to start with (log-normal and exponential, for example);
2. Calculate the expected density functions for the distributions;
3. For each sample from the data, assign it to the distribution with the highest density;
4. Formulate the loss function as the sum of mean squared error between the expected value and sample realizations;
5. Use linear optimizer to minimize the loss function to obtain the best fit given distribution shapes.

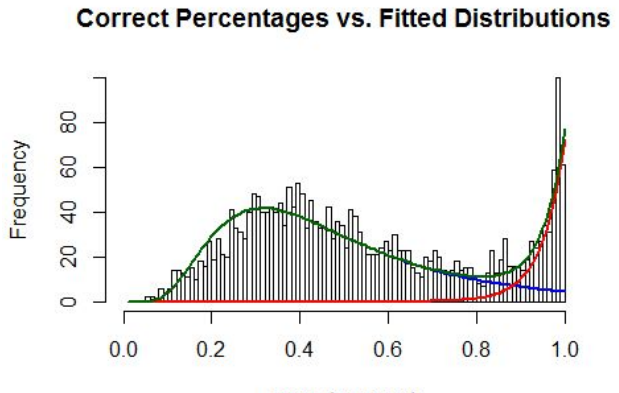
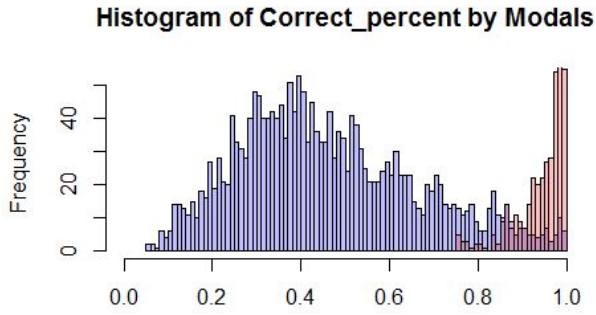


Figure 2.1.4. Blue~log-normal, red~exponential, green~joint.

The fitted distribution parameters are:

- $LogNormal(\mu_{log} = -0.867063, \sigma_{log} = 0.55014)$
- $Exponential(\lambda = 19.5126)$

Given the fitted distributions, we then assign questions to the buckets based on Bernoulli probability. As shown in figure 2.1.6A and 2.1.6B, we can nicely separate one bucket of samples from another.



Blue ~ LogNormal, Red ~ Exponential
Figure 2.1.6. Extracted modals.

Goodness-of-fit evaluation tests:

- *LogNormal*: p-value=0.1754
- *Exponential*: p-value=0.176

Therefore we may conclude that the classification is significant at 95% confidence level; there exist significant differences between

the facts/concepts that are being tested by the questions.

We investigated the player responses checking if there exists any preference over specific choice ordinal (in other words, whether the ordering of choices matter) or special pattern. The normalized distributions of percentages each choice (A, B or C) being selected is shown below.

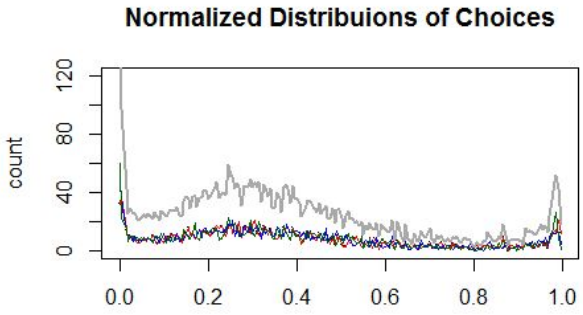


Figure 2.2.1. people are good at exluding wrong answers.

From the distribution plot above we do not observe any significant preference over any ordinal choice (in which case the preferred one would have heavier tails on the distribution plot).

We do discover that the lower-tail (near zero-percent of participants selecting any choice) has much higher density than the upper-tail. This finding may suggest that people collectively are more consistent at negating statements than reaching agreements. In the context of question-answer tests like trivia, participants behave more coherently in excluding the 'most unlikely answer' than selecting 'the most likely answer'.

As the question set covers a wide spectrum of topics, we are interested in identifying whether questions on some topics have different correct answer percentages than others. We also want to see whether category differences are correlated to the model of distribution we identified.

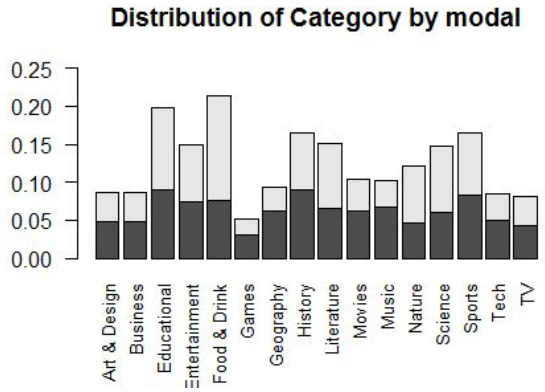


Figure 2.3.1. Dark ~ LogNormal modal, light ~ Exponential modal.

As shown in the plot above, generally there is no visible correlation between any category and distribution models. Then we shall use a linear regression to identify the effects of categories over the correct percentage.

Predictor	Estimate	Std. Error	P-value	Signif
(Intercept)	0.9535166	0.0193672	<2e-16	***
calc_modal	-0.4792319	0.010112	<2e-16	***
Business	-0.0219558	0.0244667	0.3696	
Educational	-0.0116311	0.021211	0.5835	
Entertainment	-0.0038466	0.0220793	0.8617	
Food & Drink	-0.0127952	0.0215452	0.5527	
Games	0.0135028	0.0278829	0.6282	
Geography	-0.0045641	0.0233164	0.8448	
History	**-0.043635**	**0.0214338**	**0.0419**	*
Literature	0.0117326	0.0224107	0.6007	
Movies	**-0.0517541**	**0.0231677**	**0.0256**	*
Music	-0.0388808	0.0229344	0.0901	.
Nature	0.0465838	0.0240448	0.0528	.
Science	0.0036015	0.0227282	0.8741	
Sports	0.0001722	0.0216748	0.9937	
Tech	-0.045504	0.0244147	0.0625	.
TV	-0.0087801	0.0251058	0.7266	

Table 2.3.2. Regression model showing by-category effects on percentage of correct answers with distribution controlled. R-squared=0.49.

According to the regression model, at 95% confidence level we only observe two significant effects from History and Movie questions. The model suggests that questions fall into these two categories are significantly harder with response accuracy lowered by approximately 5%.

The R-squared coefficient of the regression model is 0.49, which means only half of the variations in the correct percentages is explained. To further explore what makes people perform differently on questions, text analysis on the question and response texts is necessary.

One challenge is to find 'influential' n-grams which affect participants answering questions. The popular Term Frequency - Inverse Document Frequency (TF-IDF) approach does not apply here as some common words may actually make differences. Instead, because we have successfully classified questions into two distributions, we can compute the information gain for every phrase and rank them in the same way. This approach allows us to quickly identify phrases whose presence in the question text impacts participants' capability in finding the correct answer.

Note that information gain is unsigned, therefore the rank helps us identify subjects of influence but not necessarily the direction of the influence. To address this problem, we may calculate the slope estimates for the words in the context of a linear regression model. The estimated coefficients for the top 25 1-grams with highest information gains are shown below:

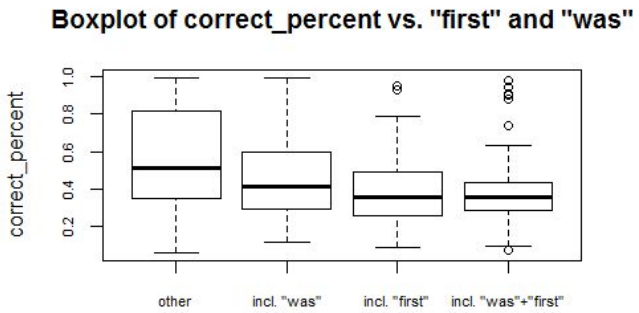
Predictor	Estimate	Std. Error	P-value	Signif
(Intercept)	0.56743	0.01763	< 2e-16	***
was	-0.04928	0.01562	0.00163	**
is	0.06813	0.0114	2.62E-09	***
what	0.0351	0.01727	0.04219	*
which	-0.02794	0.01682	0.0968	.
first	-0.1211	0.02166	2.52E-08	***
the	-0.05564	0.01011	4.12E-08	***
not	-0.06922	0.01558	9.28E-06	***
typically	**0.1425**	**0.03567**	**6.68E-05**	***
you	0.083	0.03085	0.00719	**
often	**0.13314**	**0.04937**	**0.00705**	**
countries	-0.08608	0.04017	0.03224	*
saying	0.29412	0.10248	0.00414	**
are	0.13048	0.0283	4.22E-06	***
has	-0.03757	0.02152	0.08092	.
last	-0.05273	0.04579	0.24967	
goes	0.21701	0.14525	0.13529	
want	0.22641	0.14997	0.13124	
disneys	0.30982	0.14873	0.03735	*
need	0.36552	0.14219	0.01021	*
usually	0.17101	0.05794	0.00319	**
author	-0.10197	0.0464	0.02808	*
words	-0.0868	0.04642	0.06163	.
us	-0.06895	0.02491	0.00569	**
once	-0.0553	0.04728	0.24225	
president	0.04012	0.04846	0.40783	

Table 3.2.1. Regression showing effects on percentage of correct answers from words with highest information gains. R-squared=0.13.

Table 3.2.1 (continued).

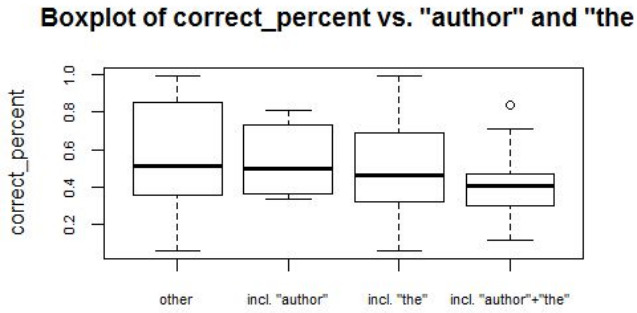
The regression model confirms the effectiveness of this information gain ranking approach as the majority of these top-ranked words displays significant influence over answer accuracy. As we can see from the list of predictors, many common words such as "was", "is", "the" are indeed top in the rank and generate significant impacts, supporting the assumption we made for not using TF-IDF.

These estimated coefficients suggest some very interesting findings. Some of them are intuitive and may be easily interpreted such as keywords like 'first' and 'was' reduce players' correct percentage, since people exhibit reducing memorization of concepts that recede farther away into the past.



if keywords are included in question_text
Figure 3.2.2A. questions in past tense are answered worse.

As appearances of keywords like 'author' and 'the', which usually refer to proper nouns like names, reduce players' correct percentage, we see that people capture general terms better than unique details.



if keywords are included in question_text
Figure 3.2.2B. questions about proper nouns are answered worse

Some others may not be very obvious. For example, we see words 'usually', 'typically', 'often' have positive relationships with correct_percent, but these words suggest some degree of ambiguity and we should expect to see more confusion and lower correct percentage in such cases. To further investigate this case, we print out the first 5 questions along with correct_percent for each of the three words.

question_text	correct%
Which U.S. road sign, usually near off-ramps, features just a single letter?	0.900063
Which of these birds is usually domesticated?	0.947559
Which of these would you NOT typically find in a blintz?	0.463089
Which of these ingredients is NOT typically used in making sourdough bread?	0.633599
In which of these is pork often replaced with turkey?	0.821382
Which of these is often squeezed to make juice?	0.984387

Table 3.2.3. Table showing sample questions containing words "usually", "typically" or "often".

From the question texts, we shall observe that, actually many questions with these keywords are "common sense" questions which can be answered without absolute an assertion of the conditions. In fact, this case demonstrates a structural paradox that seemingly unprecise questions are answered well because those that are actually unprecise would not have been made into questions.

To evaluate potential dynamics between question and available answer choices, we need to vectorize question and answer choice texts to conduct numerical analyses. Our team uses Word2Vec [17] based on popular text8 corpus to initialize the vector space, and then compute the word vector distance D between question texts and answer choices as the following method:

- Let $Q_1, Q_2, ..., Q_m$ be the vector representations of the m words in the question text;
- Let $C_1, C_2, ..., C_n$ be the vector representations of the n words in the answer choice text;

- If positive question:
  (e.g. "which is a genre of music")
  $$D = \frac{1}{n}\sum_{j=1}^{n} argmin_k \left[(Q_k - C_n)^2\right]$$
  $$for\ k = 1, 2, ..., m$$

- If negative question:
  (e.g. "which is NOT a genre of music")
  $$D = \frac{1}{n}\sum_{j=1}^{n} argmax_k \left[(Q_k - C_n)^2\right]$$
  $$for\ k = 1, 2, ..., m$$

We assume every word in the correct answer choice text is semantically related to the question, and words from the correct answer are more likely to appear in the same sentence/paragraph as the question text. For example, assume somewhere in the Wikipedia writes "jazz is a genre of music". Therefore "jazz" would have a shorter distance to "music" than "science-fiction", but not necessarily to "is", "a" or even "genre".

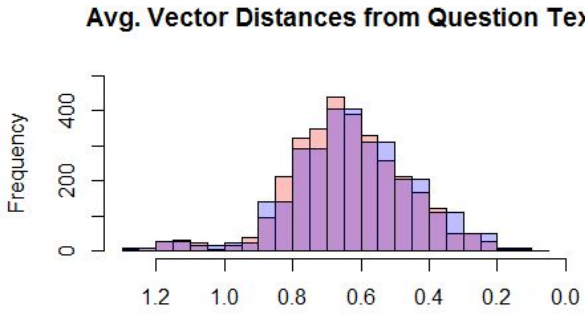


Figure 3.3.1. Correct answers are slightly closer to question text.

The median computed distance *D* between correct answers and question text is 0.6331 while the median distance between incorrect ones and question is 0.6478. Wilcoxon test produces a p-value of 0.00198 < 0.05 which indicates the difference is significant. This confirms the assumption that correct answers tend to be closer to the question text given the way we compute the word vector distances. However, the variances are large, possibly due to answer choices containing unrelated words such as "the". Our team considered enhancing this analysis by using TF-IDF scores to weight the distances per word in the choice texts but did not have time to complete due to the approaching project deadline. This would be a good start point for future researchers.

REFERENCES

[1] Collins, Harper. "Collins English Dictionary." Glasgow: HarperCollins (2004).

[2] Emma C. Boettcher. Predicting the Difficulty of Trivia Questions Using Text Features. A Master's Paper for the M.S. in I.S. degree. April, 2016.

[3] Lally, Adam, et al. "Question analysis: How Watson reads a clue." IBM Journal of Research and Development 56.3.4 (2012): 2-1.

[4] Mann, Gideon S. "A statistical method for short answer extraction." Proceedings of the workshop on Open-domain question answering-Volume 12. Association for Computational Linguistics, 2001.

[5] McCown, A. (2015, March 12). What's it like to be one of the Jeopardy! clue writers?

[6] Matt Mahoney. About the test data, 2011. URL http://mattmahoney.net/dc/textdata.

[7] Merriam-Webster, Inc. (2018). Merriam-Webster Dictionary. Retrieved from https://www.dictionaryapi.com/

[8] Salton, Gerard, and Christopher Buckley. "Term-weighting approaches in automatic text retrieval." Information processing & management 24.5 (1988): 513-523.

[9] Mahajan, D., Patil, R., Sankar, V., Word2Vec using Character n-grams. April, 2018.

[10] Fagley, N. S. "Positional response bias in multiple-choice tests of learning: Its relation to test wiseness and guessing strategy." Journal of Educational Psychology 79.1 (1987): 95.

[11] Heilman, M. (2011). Automatic factual question generation from text (Doctoral dissertation, Carnegie Mellon University). Retrieved from http://errico.srv.cs.cmu.edu/research/thesis/2011/michael_heilman.pdf.

[12] Cai, L., Zhou, G., Liu, K., & Zhao, J. (2011). Learning the Latent Topics for Question. Retrieval in Community QA. In IJCNLP (Vol. 11, pp. 273-281).

[13] Gideon Dror, Yehuda Koren, Yoelle Maarek, and Idan Szpektor. 2011. I want to answer; who has a question?: Yahoo! answers recommender system. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '11). ACM, New York, NY, USA, 1109-1117. DOI: https://doi.org/10.1145/2020408.2020582

[14] Eyecioglu, Asli and Keller, Bill (2015) ASOBEK: Twitter paraphrase identification with simple overlap features and SVMs. In: SemEval-2015: The 9th

International Workshop on Semantic Evaluation: proceedings of SemEval-2015: June 4-5, 2016, Denver, Colorado, USA. Association for Computational Linguistics (ACL), Stroudsburg, PA, pp. 64-69. ISBN 9781941643402

[15] Karampatsis, R. M. (2015). CDTDS: Predicting paraphrases in Twitter via support vector regression. Proceedings of SemEval. Pp. 75--79.

[16] Carroll, Adam Ross, "Exploring The Effects Of Multimedia Content On A Question And Answer System" (2015). All Theses. Paper 2084. 88pp. https://tigerprints.clemson.edu/cgi/viewcontent.cgi?article=3096&context=all_theses

[17] Mikolov, T., Chen, K., Corrado, G., Dean, J., Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781v3 [cs.CL] 7 Sep 2013