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

Research Proposal

Yaodong Jia, Matt Fox, Justin Melnick, Imre Patyi
Georgia Institute of Technology
Atlanta, GA 30332, USA
data@hqinsiders.com

1. Introduction

"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 exploit public data sources to study patterns in the effect of question formation, grammatical complexity, difficulty of topic, 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-up on the categories of the questions posed.

2. Project Objective
(Heilmeier Question #1)

Using 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.

3. Project Value
(Heilmeier Question #4, #5)

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, fans of trivia games will benefit from added insight on whether a question will be difficult before the answer is revealed.

We also view this project foremost as a personal learning experience which gives us a chance to explore some data, software, and techniques in a novel applied setting.

4. Survey of Related Works
(Heilmeier Question #2, #3)

Some but not many works published are specific to trivia games, mostly focusing on the question aspect of the game. Among the works many use text mining approaches to predict question difficulty [2], analyze semantic components of question [3] and estimate class/category of questions [4], while another interviews 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 in 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 Princeton WordNet [6], Merriam-Webster Dictionary [7] and grammar analysis tools such as Carnegie Mellon University's LightSIDE [8] and GloVe [9] 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.

5. Estimation of Cost and Time *(Heilmeier Question #7, #8)*

Thus far, we have spent $375 on data collection through Amazon Mechanical Turk by hiring collectors to capture data. Additionally, we will host a website to showcase our analyses, which we estimate to cost $100. We will also use Amazon Web Services and Microsoft Azure, but have used a student discount/trial for these services.

We expect to take two to three weeks for additional data collection, cleaning, and preparation. After that, we will spend another two to three weeks for data analysis, using the remainder of the semester to build out our visuals and integrate them on our website.

6. Risk Assessment *(Heilmeier Question #6)*

The most inherent risk to analyzing a "new" mobile app will be the longevity of our data feed. We are assuming that this app will increase in popularity over time and we will continue to accrue data. If the app does see a dip in daily active users, our ability to develop deeper insights will decrease.

The largest payoff is that we will be able to understand how over millions of users play a trivia game simultaneously. Historically, trivia has only been popularized in smaller segments such a Jeopardy, where three people play at a time. There has yet to be an application where millions of players answer simultaneously, and our team will be one of the first to understand how society interprets trivia questions and which unique features of the questions trigger various responses.

7. Measure of Success *(Heilmeier Question #9)*

By mid point of this project, data must be collected and exploratory data analysis should be completed.

By end of this project final models need to be built, validated, and used to discover significant patterns in the data (if any). The team is responsible of writing a final report documenting the process and findings. The master's thesis Carroll [16] suggests software and methodological tools that can be used to attack the problem of our proposal and are written at a level that we can understand without undue expenditure of effort.

8. Distribution of Work

All team members contributed equally to the writing of this proposal document. In the next phases, based on expertise and resource of each member:

- Justin will take charge of collecting, storing and cleaning data.
- Yaodong and Matt will be responsible of data analysis and visualization.
- Imre will be tasked to validate mathematical models, review project process and consolidate findings into documents.

REFERENCES

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

[2] Scott, Ralph L. "UNC-CH SILS Masters Thesis 2014-2015." North Carolina Libraries 74.1 (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? Retrieved from https://tv.avclub.com/what-s-it-like-to-be-one-of-the-jeopardy-clue-writers-1798277448

[6] Princeton University. (2015). WordNet. Retrieved from https://wordnet.princeton.edu/

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

[8] LightSIDE software, Text Mining and Machine Learning User's Manual, https://www.cs.cmu.edu/~emayfiel/LightSIDE.pdf , http://ankara.lti.cs.cmu.edu/side/

[9] Pennington, J., Socher, R., Manning, C.D. (2014). GloVe: Global Vectors for Word Representation. In: Empirical Methods in Natural Language Processing (EMNLP), 1532--1543, http://www.aclweb.org/anthology/D14-1162, Retrieved from http://nlp.stanford.edu/projects/glove/

[10] Fagley, N. S. "Positional response bias in multiple-choice tests of learning: Its relation to testwiseness 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