

NIPS 2017 Competition Proposal: Human-Computer Question Answering Competition

Jordan Boyd-Graber Hal Daumé He He Mohit Iyyer
Pedro Rodriguez

March 15, 2017

0.1 Overview of the competition

We propose to evaluate systems that can answer trivia questions that are incrementally revealed, a game called “quiz bowl”. Unlike Jeopardy or other trivia games where the players must wait until the end of the question, “quiz bowl” questions are written in a way to reward players who can answer the question earlier than their opponents.

A successful system works in two stages: they must generate candidates **and** decide when to buzz. Our approaches to solving the problem are described in the following papers.

- Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. Association for Computational Linguistics, 2015. http://www.cs.colorado.edu/~jbg/docs/2015_acl_dan.pdf
- Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. A Neural Network for Factoid Question Answering over Paragraphs. Empirical Methods in Natural Language Processing, 2014. http://www.cs.colorado.edu/~jbg/docs/2014_emnlp_qb_rnn.pdf
- Jordan Boyd-Graber, Brianna Satinoff, He He, and Hal Daumé III. Besting the Quiz Master: Crowdsourcing Incremental Classification Games. Empirical Methods in Natural Language Processing, 2012. http://www.cs.colorado.edu/~jbg/docs/qb_emnlp_2012.pdf

0.2 Keywords

Human-computer interaction, Question answering, incremental classification

0.3 Novelty

We had a similar competition as a NAACL shared task in 2016.¹ We hope to expand participation to the machine learning community. A version of the live competition (without people submitting systems) won the NIPS 2015 best demonstration award. Video of the competition is available at <https://www.youtube.com/c2kGD1EdfFw>. Our previous exhibition matches have attracted hundreds of high school students in Seattle, Dallas, and Chicago. This will open up the competition to the entire machine learning community.

1 Competition description

The basis of quiz bowl is who answers a question first. We can simulate this process asynchronously through providing words one at a time and allowing a system to choose whether to answer or not. We can compare two systems by taking each of their results and then seeing which system answered the question first.

We have constructed an API that use this protocol to query a system. Words are provided one at a time and the system can chose to submit their answer after each word the system receives.

We can use the same approach to compare humans and computers. We can replay the system performance while simultaneously reading the questions to a team of humans. The human team makes the same decision as the systems: after each word they can either answer the question or wait for more information.

We are also organizing a dual computer-human tournament to test participants question answering systems against each other and against the top human trivia masters in a separate submission. The final match between the top computer system and the top human team will be part of the live competition (if accepted).

1.1 Background and impact

This competition requires algorithms that can “think on their feet”, i.e. to incrementally process input and to decide when enough information has been received to act on those data. Successful systems require innovation in two areas: content models (to make accurate predictions, even when not all available information is available) and policies (to know when to trust the outputs of the content models—or know they won’t get better—versus waiting for more information).

This calculus is important to a number of problems. For example, *synchronous machine translation* is when a sentence is being produced one word at a time in a foreign language and we want to produce a translation in English simultaneously (i.e., with as little delay between a foreign language word and its English translation). In the machine translation setting, the content model predicts what words are going to appear in the input stream,

¹<https://sites.google.com/a/colorado.edu/2016-naacl-ws-human-computer-qa/shared-task>

even though they might not have been seen yet. This is particularly important in verb-final languages like German or Japanese, where an English translation can barely begin until the verb is seen. For the simultaneous translation problem, our content model must predict unseen elements of the sentence (e.g., the main verb in German and Japanese, or relative clauses in Japanese, or post-positions in Japanese). The job of the policy is to decide when to trust the content prediction. It must learn to balance incorrect translation versus timely translations, and must use those predictions to translate the sentence.

However, this competition focuses on a more fun task: a trivia game called quiz bowl. These questions are written so that they can be interrupted by someone who knows more about the answer; that is, harder clues are at the start of the question and easier clues are at the end of the question. Quiz bowl is a fun game with excellent opportunities for outreach, but it is also related to core applications of machine learning in natural language processing: classification (sorting inputs and making predictions), discourse (using pragmatic clues to guess what will come next), and coreference resolution (knowing which entities are discussed from oblique mentions).

1.2 Data

Training data are drawn from publicly available questions (hundreds of thousands of questions) produced by the quiz bowl community. We have previously released source code for processing these data and training models from these data.²

However, previous competitions have found that using additional external data is beneficial. Our previous shared-task winner extensively used Wikipedia data to help answer questions. We provide alignments between questions and associated Wikipedia pages to help participants use these types of resources.

The test data are provided by an independent contractor (Kurtis Droge) who has created questions that we have exclusive access to. Questions have already been created for this competition and we will commission more. After the competition, we will release the evaluation dataset.

To detect collusion, we ensure there are a small number of questions available only to a single participant; we will look for suspicious discrepancies between performance on shared vs. unique questions.

1.3 Tasks and application scenarios

The core task is answering questions incrementally. However, this task encompasses many other core NLP tasks: identifying named entities, coreference resolution.

As discussed above, this framework closely resembles the process for simultaneous interpretation, another important NLP task.

²<http://github.com/Pinafore/qb>

1.4 Metrics

We plan to use the traditional quiz bowl scoring metric. The first team to answer a question earns ten points (fifteen if the correct answer is before a designated “power” mark), but incorrect early answers are penalized by five points. Ties (which are uncommon) are broken randomly.

Manual decisions will be adjudicated using official NAQT rules.³

1.5 Baselines, code, and tutorials

In previous competitions we have provided an example python harness that uses a decision list to answer questions. This provides a simple example of how to use the code.

This year, we will also provide a bag of words neural baseline and a traditional Knesser-Ney language model baseline that achieve reasonable accuracies on the task (answers around 40% of questions correctly before the end of the question).

Our previous work describes the problem and dataset.⁴ We provide a github repository with example code and documentation⁵ and an auto-generated swagger documentation of the api.⁶

2 Organizational aspects

2.1 Protocol / Schedule

We briefly outline our competition.

1. **Summer 2017:** Advertising the system, disseminating demo/baseline system.
2. **September 1, 2017:** Server up, with development data. Participants can test their implementations against a live server.
3. **October 15, 2017:** Deadline for submitting machine entries.
4. **October 20, 2017:** Preliminary results released to teams
5. **November 1, 2017:** Deadline for submitting white papers describing participants’ systems and for electing if they want to participate in the live competition.
6. **Early December:** Live competition at NIPS

³<https://www.naqt.com/downloads/rules.pdf>

⁴Jordan Boyd-Graber, Brianna Satinoff, He He, and Hal Daumé III. Besting the Quiz Master: Crowd-sourcing Incremental Classification Games. Empirical Methods in Natural Language Processing, 2012. http://www.cs.colorado.edu/~jbg/docs/qb_emnlp_2012.pdf

⁵<https://github.com/Pinafore/qb-api>

⁶<http://petstore.swagger.io/?url=https://raw.githubusercontent.com/Pinafore/qb-api/master/swagger.json>

2.2 Rules

A complete draft of the rules is available on our task webpage.⁷ In short:

- Participants can use any dataset they would like to train models
- Participants must report results from a machine model. Humans cannot inspect the inputs or submit a human-in-the-loop system.
- Participants cannot pool access to the server to improve their results.

2.3 Competition promotion

We will promote the event on Twitter, Facebook, the HSQB forum, ml-news mailing list, our machine learning courses, and at invited talks. Both Daumé and Boyd-Graber have been discussing this line of research at invited talks and will also promote at summer conferences and on social media.

2.4 Organizing team

- **Jordan Boyd-Graber (Colorado)**: Jordans research focuses on discovering hidden structure in natural language with applications to help users sift through documents, discover when individuals control conversations, or compete against humans in games that are based in natural language. *Jordan will lead the overall organization of the competition.*
- **Hal Daumé III (Maryland)**: Hals primary research interest is in developing new learning algorithms for prototypical problems that arise in the context of language processing and artificial intelligence. This includes topics like structured prediction, domain adaptation and unsupervised learning. *Hal will lead recruitment and publicity on the East Coast.*
- **He He (Stanford)**: He develops algorithms that can reason with incomplete data in the context of natural language processing, including machine translation and question answering. *He will work on recruiting in the Bay Area, preparing data, and communicating with participants.*
- **Mohit Iyyer (Maryland)**: Mohit works on problems at the intersection of deep learning and natural language processing. He is particularly interested in designing deep neural networks for question answering. *Mohit will lead organizing human competitors.*

⁷<http://sites.google.com/view/hcqa/rules>

- **Pedro Rodriguez (Colorado):** Pedro’s research focuses on scalable learning algorithms for large, streaming data on modern computing platforms. Pedro is also the student leader of Colorado’s data science team, where he organizes internal data science competitions. *Pedro has designed the API for the competition and will lead the technical organization.*

3 Resources

3.1 Existing resources, including prizes

We will provide standard quiz bowl prizes (trophies) to the top human and computer participants.