AI Journey 2019: School Tests Solving Competition

Alexey Natekin
Data Souls, Open Data Science
Moscow, Russia
natekin@ods.ai

Peter Romov
Data Souls, Open Data Science
Vilnius, Lithuania
p@datasouls.com

Valentin Malykh
Huawei Noah's Ark Lab
Moscow, Russia
valentin.malykh@huawei.com

Abstract

Question answering is a popular complex task in Machine Learning. One challenging type of question answering is developing systems, capable to pass education exam tests. Such tests induce an intuitive evaluation metric in the same points as humans gain in exams and provide a clear set of baseline scores in terms of passing the exam. In this paper we propose a novel competition protocol that generalises exam-oriented question answering to new types of questions. We describe a shared competition task with both automatic evaluation of questions and involvement of human assessors for evaluation of computer generated essays as part of the exam tasks. We describe the data format as well as complete evaluation pipeline required for such competitions.

1 Introduction

Previously exam-oriented question answering was organized as competition on Allen AI Challenge [1] designed to answer multiple choice questions from a standardized 8th grade science exam. A continuation of the task for natural science questions was organized on a scientific AI2 Reasoning Challenge [2] as described in [3]. Recently there was significant progress in solving this task by Aristo system described in [4]. However, real exams can be more sophisticated and require not only multiple choice questions, but also more general question answering, problem solving and writing essays. In this paper we propose a competition environment that extends exam-oriented question answering to multiple types of questions as well as incorporating generic essay generation into one shared task [5]. This competition is designed to pass a school exam on Russian language with all compliance to official guidelines and an identical scoring procedure as is

2 Methods

In order to prevent potential data leakages and manual labeling exam questions, we propose a more secure competition format, where the test set is hidden from the participants. To be able to test submitted solutions in such environment we need to run solutions without the intervention of their authors. We chose Docker engine [6] to provide such isolated environments or containers. We isolate containers from the Internet, i.e. they have no Internet access from inside and no Internet addresses to access from outside. The only data which a solution has access to is a test set. In addition, a solution is able and supposed to write its output to a specified file in the file system. Each container is provided with the same amount of computational resources, including 1 NVIDIA GTX 1080 GPU, to

Preprint. Under review.

run. Participants upload their containerized solutions and submit only a code to run their solutions. Participants are allowed to use any publicly available Docker image in Docker Hub [7] thus allowing them to make solutions with any set of preferred libraries and programming languages with only limitation of 20 Gb of disk space for the whole solution.

2.1 Exam format

The exam consists of 26 questions and an essay. Each of the questions consists of text, possible attachments and answer type. Answers could be of different types: *choice* - choosing one option from the list; *multiple choice* - choosing a subset of options from the list; *order* - arranging options from the list in correct order; *matching* - correct matching of objects from two sets; *text* - answer in the form of arbitrary text.

2.2 Evaluation pipeline

Check-phase. Solution is evaluated on publicly available set of questions with known answers. This phase is important for testing solutions for potential errors and issues with evaluation system interaction. Evaluation result and system output are fully available for the participant.

Public Test. Solution is evaluated on a hidden set of questions. Results on these questions form a leaderboard during the active stage of competition. Tasks and answer options within tasks are randomly rearranged each evaluation for further defense against leaderboard probing and trying to get extra information from hidden test data.

Private Test. Solution is evaluated on the final set of questions. Results on the private test are the ones that determine competition winners. To prevent possible data-leakage the private test is created by experts using the format of official state exam.

2.3 Evaluation objectives

Each question task is evaluated by a metric which is relevant to this task type: *choice* - accuracy; *multiple choice* - union / intersection; *order* - the proportion of correctly ordered pairs *matching* - the proportion of correctly matched pairs; *text* - special evaluation function, followed by a request for human-expert assessment.

2.4 Essay evaluation

Essay evaluation comprises of two stages: automatic scoring and manual human-expert assessment. Automatic procedure evaluates basic surface-level indicators of the generated texts: no plagiarism; original topic correspondence; orthography; sentence connectivity, tautology; language errors (slang, swearing); paragraph structure; text volume (not too short/long).

Automatic scoring is given straight away and is not the final score being only a helpful utility for participants. Manual essay assessment is carried out by professional experts who follow the official grading standards of Russian state exam essays [8]. In case automatic scoring indicates that manual essay assessment would lead to 0 points, participant is informed about it and is proposed to prepare a new solution for human assessment.

3 Discussion

The proposed competition design provides a reusable framework for future competitions and environments for general question answering problems. Containerized format solves multiple common issues in competition organization: reproducible results [9, 10], secure environment with hidden test data, high flexibility in tools and approaches used by participants. Proposed data and evaluation format is suitable for many question answering problems including other knowledge domains including natural sciences and programming. This format easily extends to more sophisticated problems like Visual Question Answering [11] by simply adding relevant attachment files. And to ensure additional safety from leaderboard probing, randomizing the order of questions and answers within questions should be considered.

Acknowledgements

We thank PAO Sberbank for funding The AI Journey Challenge 2019, and appreciate all the collective effort from an expert team of this challenge from Sberbank, DataSouls and Huawei in collaboration.

References

[1] Kaggle Allen AI Science Challenge. https://www.kaggle.com/c/the-allen-ai-science-challenge.

[2] the AI2 Reasoning Challenge ARC. https://leaderboard.allenai.org/arc/submissions/public.

[3] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[4] Peter Clark, Oren Etzioni, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, Sumithra Bhakthavatsalam, et al. From'f'to'a'on the ny regents science exams: An overview of the aristo project. *arXiv preprint arXiv:1909.01958*, 2019.

[5] AI Journey 2019 competition platform. https://datasouls.com/link/ai-journey-2019-school-test.

[6] Docker platform. https://www.docker.com/products.

[7] Docker hub. https://hub.docker.com/.

[8] Criteria for essay grading on Unified State Exam 2019. https://leaderboard.allenai.org/arc/submissions/public.

[9] Rachael Tatman, Jake VanderPlas, and Sohier Dane. A practical taxonomy of reproducibility for machine learning research. *Reproducibility in Machine Learning Workshop at ICML 2018, Stockholm, Sweden*, 2018.

[10] Tatiana Likhomanenko, Alexey Rogozhnikov, Alexander Baranov, Egor Khairullin, and Andrey Ustyuzhanin. Improving reproducibility of data science experiments. *ICML 2015 AutoML Workshop*, 2015.

[11] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Farhadi Ali. Iqa: Visual question answering in interactive environments. *arXiv preprint arXiv:1712.03316*, 2017.