# Summary 6:

# Is It Smart Than an 8th Grader?

## A More Accurate Way of Judging Artificial Intelligence

Sam Diamantstein 101060342

COMP 3301, Oct 18, 2017

The performance of Artificial Intelligence (AI) models is judged using challenge problems that require the model to achieve certain goals. The Turing test(TT) was developed to determine whether an AI model can exhibit conversational behavior that is indistinguishable from a human. The TT is not a sufficient measure of AI as it demonstrates human gullibility as opposed to the AIs ability to solve problems that require complex semantic understanding.

The Allen AI Science Challenge(AI2) brought together competitors to build AI models that would answer 8th-grade science multiple choice problems. The challenge required AI models to confront contextual, factual and semantic hurdles, making it a better demonstration of AI than the more simplistic TT.

The competition was four months long and involved 170 teams from around the world. In the process of preparing the models, competitors used a training set of 2,500 questions, and a validation test set of 8,132 questions of which 800 were legitimate. In the competition, AI2 provided a test set of 2,583 questions used to produce the final test scores for each teams AI model.

The competition used four-way multiple choice questions resulting in a baseline score of 25% using random answers.Through using Lucene search, an information retrieval engine, a 40.2% score on the test was achieved. The top three contestants in the AI2 challenge all used information retrieval based features, that searched for answers in a body of information like Wikipedia. According to Chaim Linhart, the first place contestant,scores as high as 55% could be achieved with information retrieval methods alone.The three winners were Linhart at 59.31%, Benedikt Wilbertz in second at 58.34% and Alejandro Mosquera at 58.26%.

Linharts winning approach predicted the correctness of each possible answer by using 15 small gradient boosting models. The models were divided into two categories, one using information retrieval using weightings and stemmed words to optimize performance.The other category assessed the qualities of the questions and answers, looking at length and type.Linharts AI model, composed

of many smaller models, enabled it to learn about many features of the questions and answers. Linhart's approach of using many models was crucial given the small amount of training data in the competition.

Wilbertzs model used information retrieval systems on a larger scale relative to his competitors. Wilbertz also used a string hashing feature where a word and its definition were hashed, thereby training the model which pairs were correct.

In third place, Mosquera used a three-way classification system, turning four answers (a,b,c,d) to into 12 pairs of answers( A,C),.,(D,C) and then grouping the 12 into 3 categories. The approach made it easier for supervised learning algorithms to make a judgment on the correctness of an answer because of a relative ranking of the choices. Mosquera also used information retrieval features based on scores from Elastic and Lucene search.

The AI2 challenge demonstrated that AI models were unable to achieve a deeper understanding of the meaning underlying each question. In addition, there were problems associated with using reasoning to pick the appropriate answer. The AI never surpassed a 60% score, demonstrating the failure to achieve a semantic level of meaning that entails true AI.