

CSCI 5622: Final Project Proposal

Project Title:

Teaching Machines Reading Comprehension

Group Members:

Avinash Dhulehole, Prajnaya Satish, Shreya Verma, Vikhyat Goyal

Abstract:

Inference is a fundamental skill that a human develops when he performs reading comprehension problems. The challenge is not only to identify the relevant portion of a text but also to extract the inferred context. This is what represents intelligence and is a necessary and important area to cover when trying to develop artificial intelligence. A machine tackling reading comprehension may be a step in the direction of artificial intelligence.

Traditional approaches to automatic reading comprehension are rigid as they depend on a set of handcrafted rules that use lexical and syntactic patterns to extract the right answer from a paragraph. Language itself is ambiguous and rules are never sufficient to make decisions. Machine learning may be a solution here because the machine automatically computes the rules without any interference from humans.

Dataset:

In this project, we tackle Reading Comprehension using [SciQ Dataset](#) released by Allen Institute for Artificial Intelligence.

Approach:

Representing language as mathematical models is a challenge and there are very few machine learning models that can help analyze it. For this project, we have decided to use a neural network architecture. This will give us the required depth in analysis to overcome the challenges which come with inherent ambiguity of language. In particular, we plan to use LSTM. Our dataset comprises of a question and a set of possible answers for a support paragraph.

Fossil records support the idea that periods of little change may be interrupted by bursts of rapid change, an evolutionary model contrasted by what?

Correct Answer	Distractor 1	Distractor 2	Distractor 2
gradualism	elitism	darwinian	minimalism

Support: When geologic and climatic conditions are changing, evolution may occur more quickly. Thus, long periods of little change may be interrupted by bursts of rapid change. This model of the timing of evolution is called punctuated equilibrium . It is better supported by the fossil record than is gradualism.

Typically, a reading comprehension task consists of a long paragraph (long sequence of words) followed by a question which calls for retention of memory. One of the advantages of LSTMs is they are capable of retaining memory over very long span.

One of the key challenges we might face in this project is embedding real world context into a machine. For example, the question in the above sample asks for *“an evolutionary model contrasted by”* but the support paragraph says *“... is better supported by the fossil record than is gradualism.”* Clearly, a lot of linguistic tools such as synonym, context and paraphrasing are used here to match the two pieces of text. Hopefully, the LSTM model is able to encode this information while trying to extract the answers from the paragraph in scope.