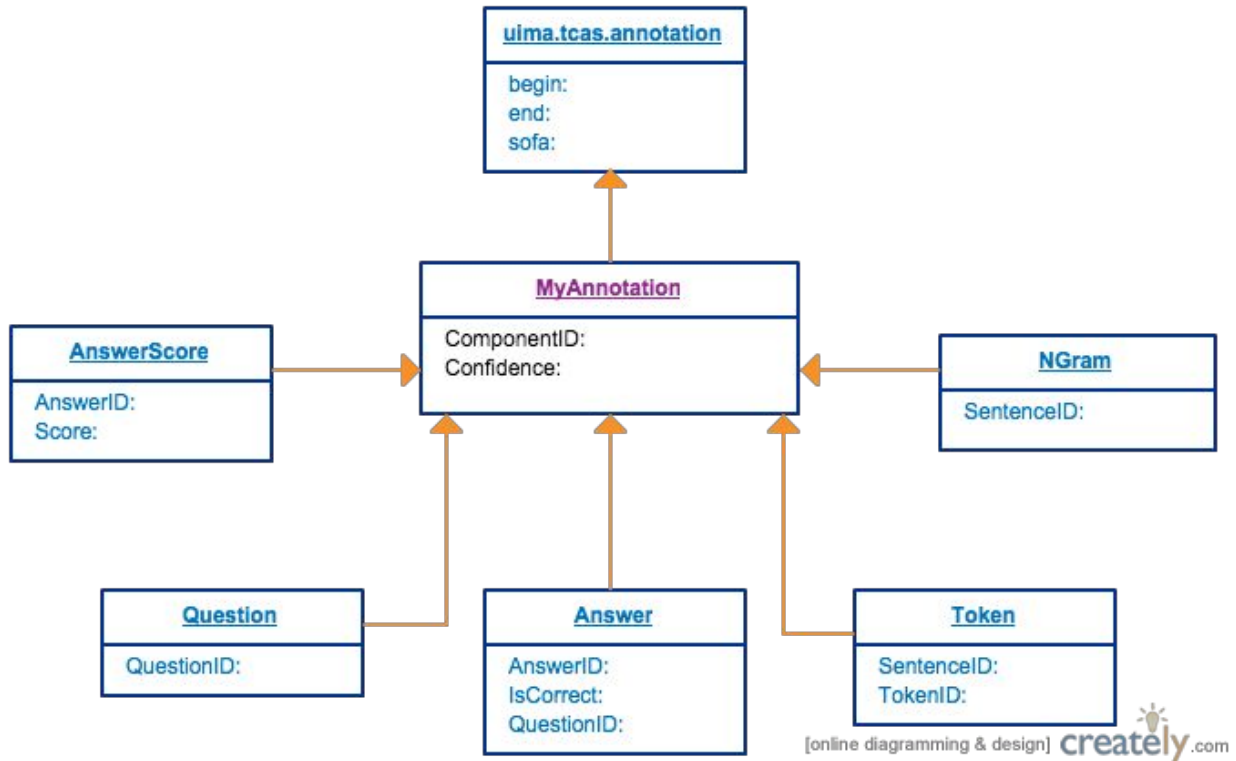# Report PI2
Yash Gaur (ygaur@andrew.cmu.edu)

Below is the **UML diagram** of the type system followed in this project:

Explanation: Since every annotation needs to have name and confidence of the component that produced it, we make a superclass called **MyAnnotation**, which contains these fields. Every other information object will inherit from this class. MyAnnotation will inturn inherit from **uima.tcas.annotation**.

The input files are fed to '**TestElementAnnotation**', which will generate 2 kinds of information objects: **Question**s and **Answer**s. Question is an annotation which stores the question. It also stores with it a QuestionID which uniquely characterizes this question, in the case when there are multiple questions on the page. Answer is an annotation which stores an answer candidate. It will have a unique identifier called AnswerID, a field called IsCorrect, which will tell if the answer is a correct answer to the question and also a field for storing the QuestionID.

Note:*This design will also work if there are multiple questions and answer sets in an input file*.

The Question and Answer annotation are then fed to the '**Token Annotation'**, which produces **token**s. Each token will contain a sentenceID and TokenID. SentenceID will store either the AnswerID and QuestionID of the sentence from which the token is coming. TokenID is the

sequence number which will be used to calculate the adjacency of tokens while calculating NGrams. The combination of SentenceID and TokenID uniquely determines any token in the pipe.

These tokens are then fed to '**NGram Annotation**', which produces **NGram** information objects. Each NGram will also contain the SentenceID depicting where that NGram comes from.

These NGrams are then fed to the '**Answer Scoring**', which will calculate **AnswerScore** objects for each answer. These objects will contain the score and the AnswerID. Score contains the a float value which will tell you the likelihood of that answer being the correct answer and AnswerID will contain the identifier of the Answer for which the score is being calculated.