Mario Piergallini
mpiergal

The first type is the **BaseType** which extends the **BaseAnnotation** type by adding features for the source of the annotation and confidence. The **Token** type is just an extension of the **BaseType** and does contain any additional features since the beginning and end points are included in the **BaseAnnotation**. The **Sentence** type has the list of tokens comprising the sentence and lists of the unigrams, bigrams and trigrams in the sentence. The unigrams, bigrams and trigrams are represented by the type **NGram** which contains references to the tokens in the n-gram and the length of the n-gram.

It would seem like the **Token** type and unigrams would be redundant, but depending on how you're using unigrams they might be different. For example, you could just be recording the presence of an n-gram, rather than wanting to have references to each token of that n-gram in the sentence. For unigrams, multiple occurrences of the same word would only be represented as one unigram. In that case, you might want to add another feature to keep track of all the locations of a particular n-gram.

The questions and answers both inherit from the **Sentence** type. The **Question** type contains a reference to the potential answers associated with it and the evaluation of the system's performance on that question. The **Answer** type has a feature isCorrect which is just the value of the gold standard annotation for that answer and a reference to the **AnswerScore** (a value between 0 and 1) generated by the system for that answer. The **Evaluation** type simply has the precision value, which would be generated from the answer scores and gold standard values.