Hw1 Documentation, LTI 11791

Abhimanu Kumar
Andre Id - ABHIMANK

September 2013

1 Class Description

I describe here the design considerations, UIMA types issues and other NLP issues that I took into account while designing the type system for answer evaluator problem in HW 1.

Token: start, end, (both are covered by annotation class) stringText, ascii(yes/no), tokenId, pos, lexicalFeatures(morphology) Ngram: start, end, tokenList, ngramId, n, semanticFeature(role) Sentence: language, tokenlist, length, ngramlist, sentenceId, start, end, parseTree (built in type) Question: sentence, questionId, answerList Answer: sentence, correct(yes/no), score, rank, answerId, questionId ConfidenceEntity: source, confidence Evaluator: questionId, precisionAtNlist, n,

1.1 Class Token

This class is the basic building block token class that models every token encountered in the dataset. It extends the UIMA's Annotation class thus it doesnt need a start and end member. Its stringText member is the actual string of the token. The ascii is boolean that store whether there are any non-ascii charchter in the token. The tokenId is the unique token id assigned and **partOfSpeech** is the pos tag for the token. A special NLP attribute called **lexicalFeatureList** is provided that will be helful in deciding the right answer score. It includes token features such as its morphology etc.

1.2 Class Ngram

This is the class that encapsulates the group of tokens that are taken as bi-gram tri-gram etc. Besides the normal attributes (length, tokenList) that I included, I provided a semantic member called **semanticFeatureList** that includes ngram attributes such as **semantic roles** etc.

1.3 Class Sentence

This class includes the **language** attribute of the sentence since this is the appropriate class where language can be defined. besides the normal attributes (sentenceId, tokenList, lengthm ngramList) it has semantic member **parseTree** that helps in deciding the answer score.

1.4 Class Question

This class encapsulates the sentence body of the question, the probable list of answers that it is assigned and a uniqueid. Note ythat we do not store answers as member in this class but just the answerId.

1.5 Class Answer

This class encapsulates the answers provided to a question. Besides the unique id (answerId) and sentence attributes it also stores the questionId to which it is answer to. It also has **rank** member that gives the rabk of the answer and **score** that gives the correctness score. It also has a boolean attribute **correct** which just provides whether the answer sentence in general is correct or not; this helps in easier ranking.

1.6 Class ConfidenceEntity

This is the class that encapsulates the confidence score of the annotation and the source of the annotation

2 Special Points

1) Almost all classes extend Annotation class except Answer. Answer class extends ConfidenceEntity class which encapsulates confidence and source of the answer annotator. 2) semanticFeatureLists, lexicalFeatureList and parseTree are FSlists and their type is TOP. 3) The Token class is a member instance of Ngram class. The Ngram and Token class are member instance of Sentence class. The Sentence class is the member instance of Question and Answer classes.