

Project Individual 2: UIMA Type System

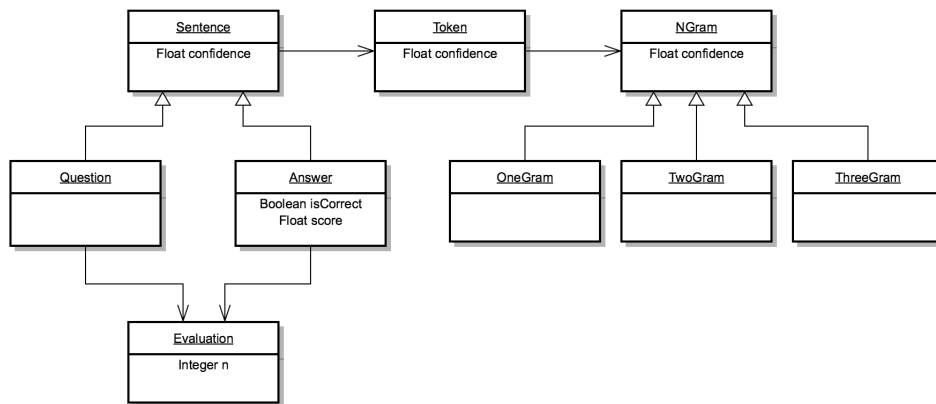
Leah Nicolich-Henkin

September 14, 2015

Overview

This report outlines the UIMA type system for a task which evaluates whether answers to a question are correct. The system involves a series of annotators, which read the input file and decompose it on different levels. After annotations are complete, it ranks the answers according to how well they match the question, chooses the top N , where N is the number that are actually correct, and then evaluates the output.

Type System



The type system includes three main types of Annotations: the Sentence, the Token, and the NGram, each of which also has a confidence feature. There are two types of Sentences: Question, and Answer. When the input is first read by the system, it is annotated as one Question and multiple Answers. The Sentences are then annotated with Tokens, which in turn are used to annotate the N-Grams in the Sentence. The NGram Annotation is extended by three different Annotations: OneGram, TwoGram, and ThreeGram. Both Questions and Answers are annotated with tokens, but Answers additionally have a Boolean feature **isCorrect**, determined based on the input file, and a Float feature **score**, which is calculated with reference to the NGram Annotations of both the Answer itself and the associated Question.

Once the input has been fully annotated and the scores calculated, the Evaluation component can compare the scores of the different Answers, rank them, choose the top N as correct, and then compare them to their **isCorrect** values.

Methodology

The organizing principle to this type system is that Question and Answer, by extending a Sentence annotation, are treated as essentially the same type, with the same annotations. This makes it easy to view them as analogous types and to compare them. Similarly, the Annotations used for comparison are all different NGrams, and can, by and large, be manipulated and compared in the same ways. The input flows through the different annotations, being broken into increasingly small and organized pieces, before those pieces are incorporated to evaluate the sentences.

An alternative structure might have used a type system based on Strings, rather than Annotations. However, in keeping with the UIMA system of Annotations, we can assure that the reference points of the types remain consistent throughout the system.