Logical Data Model and UIMA Type System Design & Implementation Report

Chenying Hou
Andrew ID:chenyinh

September 11, 2013

1 Introduction

My logical data model for a sample information processing task aims to accomplish the five phases-pipeline processing task. The main motivation of my design is that the annotations can be used to realize the processing pipeline from test element annotation, token annotation, NGram annotation to answer scoring and the final evaluation. Also based on the basic function of the pipeline, I try to extract more characteristics from the raw information text(which are original Questions and Answers) and add more features in the annotations.

Considering the input of this system,the basic logical data type should be the Question and the Answer type. These two types are associated with a span having begin and end which can locate the annotation.

We also should consider the medium types needed during processing. The Token annotation part will generate tokens from the Question and Answers. So we need to build a new type to record that information. Also the Token annotation is bound with a smaller span which also needs begin and end features. Further, in order to retrieve more feature of the token, we can record whether this token is a noun, a verb or something else. This part of speech information will help in parsing the sentence more correctly.

Then in the NGram Annotation, we need a type to record the result of 1-, 2-, 3-grams consecutive tokens. And a NGram type should contain what tokens this type contains.

Next step, the system will use some method to score the answer and compute a final score for each answer. Since each score are bounded with an answer and an answer has multiple scores, we record the score information within answer type. Also, different score are made by different processor/author, we add that information to each score. In order to quickly find a score for certain answer

and have a clear view of how different the scoring methods are for the same answer, we put this score information within the answer type.

Therefore, this type system includes five annotations, which is **Answer**, **Score**, **Question**, **NGram**, and **Token**. They will be introduced in detail in next section.

2 Annotations Design

There are 5 main types in this system: Answer, Question, Token, Score and NGram. They all in the "model" name space in my package. All types have two same features called "casProcessorId" and "confidence". These two fields indicate where this annotation was originally made by (by which processor or phrases in the pipeline),and how confidence the annotation was. I will not explain them in the following report.

All the annotations are inherited from a CAS base annotation type called "uima.tcas.Annotation", except for Score type. The Score type is inherited from TOP based annotation because it does not need "begin" and "end" feature which will be mentioned below.

2.1 Question

This type refers to the input questions. It has following features:

begin : default feature extended from base annotation type and it indicates the beginning of the span of Question annotation.

end : default feature extended from base annotation type and it indicates the ending of the span of Question annotation.

casProcessorId

confidence

2.2 Answer

This type refers to the input answers. It has following features:

begin : default feature extended from base annotation type and it indicates the beginning of the span of Answer annotation.

end : default feature extended from base annotation type and it indicates the ending of the span of Answer annotation.

casProcessorId

confidence

isCorrect : a boolean type feature, shows whether this answer is correct or not.

scoreList : a FSArray structure and its elements are **score** type made by different evaluation methods/authors. The **score** type is integrated within answer type because one specific answer must related with many scores. Different scores have different author and different value which is defined in the **score** type.

2.3 Token

This type refers to the element in the question and answer sentences. It has following features:

begin : default feature extended from base annotation type and it indicates the beginning of the span of this token annotation.

end : default feature extended from base annotation type and it indicates the ending of the span of this token annotation.

casProcessorId

confidence

wordClass : indicates the part of speech of a token, such as whether this token is a noun a verb or others which will help in further parse the sentence and match the token in dictionary.

2.4 NGram

This type records the data needed for NGram processing. It has following features:

begin : default feature extended from base annotation type and it indicates the beginning of the span of this NGram annotation.

end : default feature extended from base annotation type and it indicates the ending of the span of this NGram annotation.

casProcessorId

confidence

elementType : the type of elements in this NGram, such as a token.

elements : a FSArray structure, recording the element in this NGram.

2.5 Score

This type is inherited from TOP. Since score must be bound with an answer in my design so it does not need the "begin" and "end" features. Different algorithm/author/ can computed different scores, so I record both casProcessorId information and corresponding score information. It has following features:

casProcessorId

confidence

score : a double structure to record precise score value.