Software Specification

for

NTCIR QALab CMU Baseline

Version 1.0

Prepared by Di Wang

Language Technologies Institute
Carnegie Mellon University
School of Computer Science

Contents

1 Introduction

This document describes an UIMA based modular question answering (QA) pipeline that automatically answers multiple-choice questions for the entrance exams about world history, which provides an end-to-end baseline system for NTCIR QALab-1 challenge (stage 1). The pipeline consists of a XML collection reader, a question analysis annotator, a document retrieval based evidence collector, a rule based answer selection, and evaluation CAS consumers. This baseline system can correctly answer up to 53 of all 143 questions from the 1997, 2001, 2005, 2009 training datasets provided. To facilitate future collaborative efforts for designing and implementing question answering systems for the world history exams, the type system, collection reader, QA phases, and evaluator can also be served as a modular software platform for evaluating component performance.

Given a topic, contextual information (a short excerpt on the topic), and specific question instructions, the question analysis component generates verifiable assertions for each answer choice. Because we focus mostly on information retrieval-based answer selection, these assertions are converted into search queries. Then we validate assertions by running the queries against an indexed collection such as Wikipedia. The most plausible answer choice is selected based on retrieval scores of found documents.

The rest of the document is organized as follows. In Section 2, we outline the overall architecture of our QA pipeline. The UIMA type system of this baseline is discussed in Section 3. In Section 4, we describe each phase and its baseline module in more details.

2 System Architecture

The Apache UIMA[1] is one of the state of the art frameworks for natural language processing [1]. It is a powerful open-source tool that provides capabilities for serialization, advanced flow control, and distributed processing. This is why, after analyzing the task, we decided to create an easy-to-extend UIMA-based system that contains question analysis, document retrieval based evidencing, and answer selection modules. Figure 1 shows the pipeline phases overview of our baseline system.

3 Type Systems

Our UIMA Type System represents both 1) topics and their metadata parsed from test XML document and 2) intermediate data that used between phases.

Several important types about topics and their metadata is shown in Figure 2. Each TestDocument contains multiple topics (QuestionAnswerSet). And QuestionAnswerSet contains a Question and a list of AnswerChoice. Metadata about

[1] `http://uima.apache.org/`

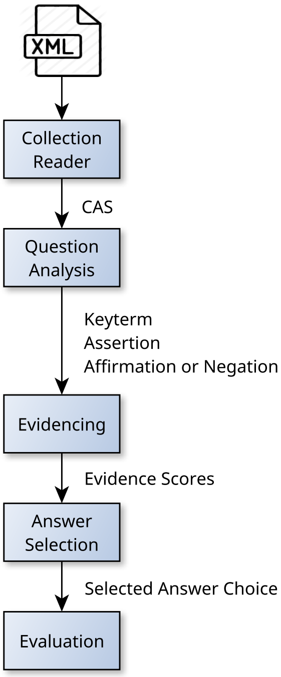

Figure 1: Phases Overview

a question and answer choice is stored as features such as question type and choice number.

Major types to store intermediate data in our baseline is highlighted in Figure 3. Basically, our question analysis module generates AnalyzedAnswerChoice and Assertions, and then evidencing module assign AssertionScores. Finally the answer selection module will make its choice based on AssertionScores.

4 Pipeline Phases

4.1 Input Collection Reader

The test document collection reader parses the information from the input XML files containing topics, and stores them as annotations in UIMA CASs that can be processed by UIMA pipelines easily. Optionally, the collection reader can be configured to load gold standard files for developing and training purposes.

The baseline repository also includes a XMI collection reader that can take saved XMI (serialized CASs) as input.

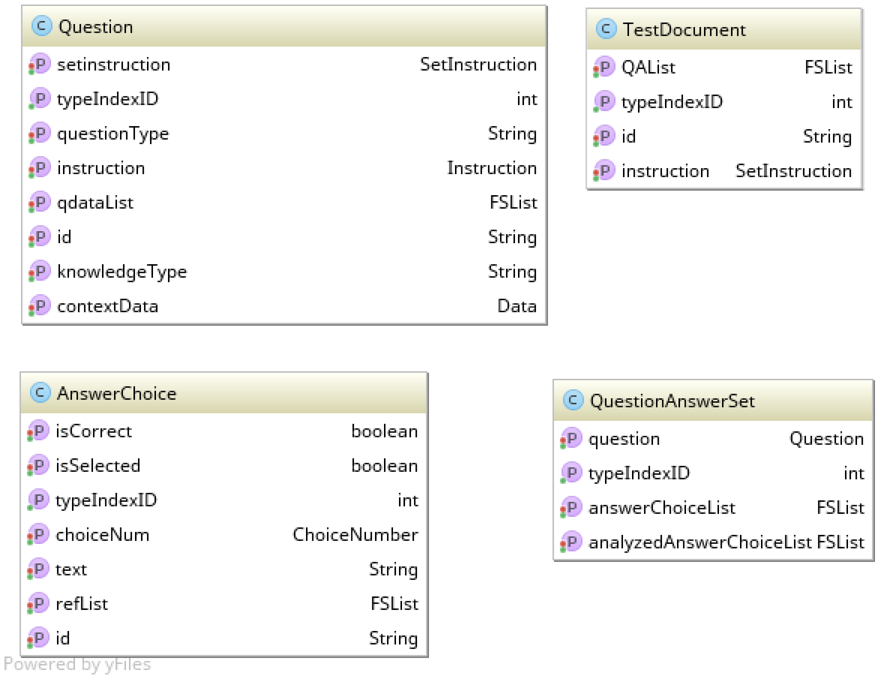

Figure 2: Important types about topics and their metadata

4.2 Question Analysis: Hypothesis Generation

Our question analysis module analyzes the question provided in test XML document along with its constituents such as higher level instruction, section text, question-specific instruction, question-specific text and references associated with question and answer choices. This component generates hypothesis and its assertions which are validated in subsequent modules.

After carefully examining the NTCIR-11 QALab task and the XML data about entrance exam provided to us, we devised our strategy to build generic framework for answering such multiple choice questions. The data mostly contained questions that ask user to identify whether given statment in answer choice is correct or incorrect. Other questions were fill-in-the-blank type, where user has to choose correct word(s) from available choices and fit in the blank to make the sentence correct in given context. Some questions involve image analysis which we are ignoring in our work and choose not to answer. So, the overall task is reduced to just validate wether given sentence in answer choice or sentence formed by replacing answer choice in fill-in-the-blank sentence is relatively correct or incorrect with respect to other possible answer choices. We call such sentences as *assertions*. Depending on whether we are evaluating assertion correctness or incorrectness, it can take boolean value true/false. Each (question,answerChoice) forms a *hypothesis*. As each answer choice many times requires multiple phenomena to be true simulteneously, each hypothesis can have multiple assertions. Hypothesis object is built by extracting all relevant information. We have tried

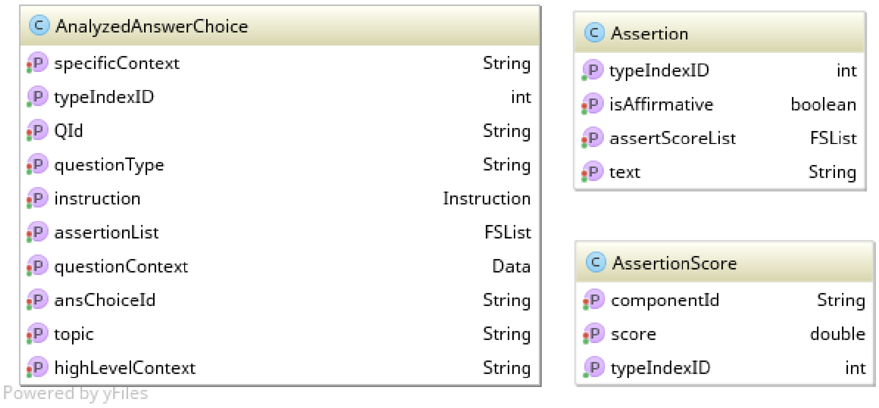

Figure 3: Important types about intermediate data

to make hypothesis as self-sufficient as possible so that subsequent components need to analyse only input hypothesis. The examples of some hypothesis objects are shown in Table 1 and Table 2.

4.3 Evidencing: Document Retrieval

Each hypothesis generated during question analysis has one or more assertions attached to it. To evaluate the soundness of an assertion, we create a retrieval query from the assertion, ran the query, obtain the scores of one or more highly ranked documents, and generates a evidence score for the assertion.

The document retrieval module in the baseline system tries to query a Solr indexed corpus of Wikipedia articles. It outputs the relevance score of top ranked document and number of retrieved documents. In our baseline, we trained a logistic regression model on number of search results and relevance score of top document and determined the weights of each of these features. Then, we compute the final score as follows:

$$EvidenceScore(A) = w_0 * numSearchResults(A) + w_1 * relevanceScore \quad (1)$$

The intuition is that the assertion with the highest score is likely to be the correct one. If the question asks which assertion is **incorrect**, we prefer the assertion with the lowest score.

4.4 Answer Selection

The answer selection module in our baseline makes decision based on the evidence scores for each answer choice's assertions from all evidencing components. All evidence scores of a answer choice will be summed up to be combined as final evidence score. If the question ask for which answer choice is correct, the

QId	Q2
AnsChoiceId	(1)
QuestionType	sentence
Topic	the history of occupations and labor
High Level Context	Writing about trends among highly-educated people during the Ming period, the Qing period scholar Zhao Yi states that from the (1)Tang and Song periods onwards, most of those who excelled in culture and the arts were those who had passed the Imperial examinations, but in the (2)Ming period, there was a shift toward figures outside the bureaucracy. . . . bureaucracy emerged after culture matured in cities, due to (3)the development of commerce and industry, focused mainly on the Jiangnan region, with pictures and publications coming to possess wide-ranging value as products.
Specific Context	Tang and Song periods onwards, most of those who excelled in culture and the arts were those who had passed the Imperial examinations
Assertion	Ouyang Xiu and Su Shi are writers representative of the Tang period. [true]

Table 1: Hypothesis- Example 1

answer choice with highest final evidence score will be selected, otherwise the lowest final evidence scored answer choice will be chosen.

The baseline's simple answer selection module also includes evaluation functionalities that prints out final accuracy and final scores of selected answer choice.

References

[1] H. Gómez-Adorno, D. Pinto, and D. V. Ayala. A question answering system for reading comprehension tests. In *MCPR*, pages 354–363, 2013.

QId	Q4
AnsChoiceId	(2)
QuestionType	symbol-term
Topic	the history of occupations and labor
High Level Context	Writing about trends among highly-educated people during the Ming period, the Qing period scholar Zhao Yi states that from the (1)Tang and Song periods onwards, most of those who excelled in culture and the arts were those who had passed the Imperial examinations, but in the (2)Ming period, there was a shift toward figures outside the bureaucracy. . . . From the middle to the late Ming period, a succession of artists and writers outside the bureaucracy emerged after culture matured in cities, due to (3)the development of commerce and industry, focused mainly on the Jiangnan region, with pictures and publications coming to possess wide-ranging value as products.
Specific Context	the development of commerce and industry, focused mainly on the Jiangnan region
Assertion(1)	From the middle of the Ming period, while handicraft industries such as silk and cotton textiles developed along the lower reaches of the Yangtze River, grain-producing regions spread along its middle reaches, and the expression ”Huguang shou, tian xia zu (if Huguang ripens, all is well)” emerged [false]
Assertion(2)	Shanxi merchants and Xin'an merchants flourished, and mutual aid organizations called kongsi (clan hlls) and Zujie (concession) were established in cities in each region. [true]

Table 2: Hypothesis- Example 2