

11-791 Design and Engineering of Intelligent Information System Fall 2014 Project

Building a Pipeline for Biomedical Question Answering

Important dates

- **Hand out: November 22.^a**
- **Milestone 3 (M3) : Exact answer generation. Turn in: December 1.**
 - You are required to submit all the sources for your components via a Maven release like you did for previous assignments.
 - We expect you properly accomplished all the milestones, reported the evaluation results of your final system, and wrote a final summary of your findings.
 - Name your presentation slides as `project-teamID.pptb` and put it in the `src/main/resources/docs`.

^aThis version was built on November 22, 2014

^bOr other formats, e.g., pptx, odp, pdf, etc.

Submission Structure:

You should always organize your project in the same hierarchy as shown below for all your submissions:

```
project-teamYYY-XX
|- pom.xml
`- src
    `- main
        |- java/
        |   |- **/*.java          /* your Java code goes here */
        |
        `-- resources
            |- input              /* you should expect input from this directory */
            |- CpeDescriptor.xml  /* the entry point for the example */
            `-- bioasq
                |- collection      /* descriptors for collection reader,
                |- type            * example typesystem
                |- concepts        * example concept extractors,
                |- documents       * example document retrievers,
                |- triple          * example triple extractors */
```

Several notes about organizing your Maven project and other additional information:

1. **Submission:** The same way as you did for Homework 0 and 1, etc. (set up GitHub repo, create Maven project, write your code, submit to Maven repo), except that the name has changed to project-teamYYY-XX (XX is a two-digit number from 01 to 07, and YYY is the last three digits of your course section, e.g., 791).
2. **Your source files and descriptors:** `java/` means you should create directories (or packages in terms of Java program structure) iteratively from `java`.
3. **Comments, Javadocs, and documentations:** They are still important if you want us and other users to better understand your code.

Useful information

1. Please visit Piazza regularly to check if a newer version is published. We may have new versions or revised instruction at the beginning of each milestone.

We expect that most of the general communication between the instructor team and students will take place on Piazza <https://piazza.com/class/hyvsubeilei6dd>. For private questions, e.g., regarding grades, you may contact instructors by e-mail. Your friendly TAs are: Avner Maiberg (amaiberg@andrew.cmu.edu), Parag Argawal (paraga@andrew.cmu.edu), Leonid (Leo) Boytsov (srchvrs@cmu.edu), and Xuezi (Manfred) Zhang (xueziz@andrew.cmu.edu).

2. Again, both source files and pdf file of this assignment are publicly available on GitHub

<http://github.com/amaiberg/software-engineering-preliminary>

Please feel free to fork the project and send a pull request back to me as some of you did for Homework 0 for any error. Or you can just report an issue at

<http://github.com/amaiberg/software-engineering-preliminary/issues>

Milestone 3

Completing your BioASQ pipeline

In this task you will implement components to retrieve exact answers for a biomedical question answering pipeline (based on BioASQ's Task1b Phase A). We advise you to focus on either yes/no, factoid, or list questions. For those who are still interested in summary questions, please contact the TAs. You will create your own pipeline based on a provided archetype framework from M1. Note: the **only requirement** is that you use the typesystem provided in the archetype. You are otherwise permitted and encouraged to use any additional types and source code you may wish.

The final milestone will be done in the spirit of a friendly competition. The team with the best and/or most creative pipelines will be given a yet-to-be-specified award. Note, representatives from Roche Pharmaceuticals will also review the results of the most promising teams. For those who have done well in this task and are interested in internship/research opportunities in the biomedical domain, please contact Avner Maiberg or Zi Yang (ziy@cs.cmu.edu).

M 3.1 Types and evaluation metrics

The following are the output types described in the official BioASQ evaluation documentation¹ [1]:

Yes/no Answers

For each yes/no question, the 'exact' answer of each participating system will have to be either 'yes' or 'no'. The response will be compared against the golden 'exact' answer (again 'yes' or 'no') that the *BioASQ* team of biomedical experts will have associated with the question.

Example: Here the JSON key/value pair `exact_answer` corresponds to a yes/no answer.

```
"body": "Are there any DNMT3 proteins present in plants?",  
"concepts": [
```

¹Please note that synonyms of correct answers for factoid and list questions are also acceptable. You should do a synonym expansion of the golden set provided to get a true evaluation of your system

```

    "http://www.uniprot.org/uniprot/CMT1_ARATH",
    "http://www.uniprot.org/uniprot/CMT2_ARATH",
    "http://www.uniprot.org/uniprot/CMT3_ARATH"
  ],
  "documents": [
    "http://www.ncbi.nlm.nih.gov/pubmed/21150311",
    "http://www.ncbi.nlm.nih.gov/pubmed/21060858",
    "http://www.ncbi.nlm.nih.gov/pubmed/10781108",
  ],
  ...
  "exact_answer": "Yes.",
  "type": "yesno"

```

Factoid Answers

For each factoid question, each participating system will have to return a list of up to 5 entity names, ordered by decreasing confidence. The *BioASQ* team of biomedical experts will have associated with each factoid question a single golden entity name, as well as possible synonyms of that name.

Example: Here the JSON key/value pair `exact_answer` corresponds to a factoid answer.

```

"body": "Is Rheumatoid Arthritis more common in men or women?",
"concepts": [
  "http://www.disease-ontology.org/api/metadata/DOID:7148",
  "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&...",
  "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&...",
  "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&...",
  "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&..."
],
"documents": [
  "http://www.ncbi.nlm.nih.gov/pubmed/23217568",
  "http://www.ncbi.nlm.nih.gov/pubmed/22853635",
  "http://www.ncbi.nlm.nih.gov/pubmed/1563036"
],
...
"exact_answer": [
  "Women"
],
"id": "5118dd1305c10fae75000001",
"type": "factoid"

```

List Answers

For each list question, each participating system will have to return a list of entity names, jointly taken to constitute a single answer (e.g., the most common symptoms of a disease); for practical purposes, the maximum

allowed size of each returned list may be limited (e.g., up to 100 names, each one up to 100 characters).

Example: Here the JSON key/value pair `exact_answer` corresponds to list answers.

```
"body": "What is the most prominent sequence consensus for the polyadenylation si",
"concepts": [
  "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&",
  "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&",
  "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&",
  "http://www.nlm.nih.gov/cgi/mesh/2012/MB_cgi?field=uid&exact=Find+Exact+Term&"
],
"documents": [
  "http://www.ncbi.nlm.nih.gov/pubmed/1915889",
  "http://www.ncbi.nlm.nih.gov/pubmed/1993703",
  "http://www.ncbi.nlm.nih.gov/pubmed/6194440",
  "http://www.ncbi.nlm.nih.gov/pubmed/7901430",
  "http://www.ncbi.nlm.nih.gov/pubmed/1712333",
  "http://www.ncbi.nlm.nih.gov/pubmed/2513486"
],
"exact_answer": [
  [
    "AATAAA"
  ],
  [
    "AAUAAA"
  ]
],
...
"type": "list"
```

Evaluation

Please refer to table 3.1 to see the evaluation used for each type.

Question type	Participant response	Evaluation measures
yes/no	'yes' or 'no'	accuracy
factoid	up to 5 entity names	strict and lenient accuracy, MRR
list	a list of entity names	mean percision, recall, F-measure

Figure 3.1: Evaluation metrics

Also, please review the definitions of the evaluation metrics.

- Accuracy:

$$Acc = \frac{c}{n}$$

(where c is the number of correct answers and n is the number of questions)

- Strict and Lenient Accuracy:

$$SAcc = \frac{c_1}{n}$$

$$LAcc = \frac{c_5}{n}$$

Where,

- c_1 is the number of factoid questions that have been answered correctly when only the first element of each list is considered.
- c_5 is the number of factoid questions that have been answered correctly when all the elements of the list are considered (i.e., the list contains at least one synonym of the golden entity name).

- MRR:

$$MRR = \frac{1}{n} \sum_{i=1}^n \frac{1}{r(i)}$$

Where $r(i)$ is the topmost position of the element containing the golden entity name (or its synonyms). If no element is found, then $r(i) = +\infty$, or $MRR = 0$.

- Precision and Recall:

$$P = \frac{TP}{TP + FP} \quad (\text{where TP are true positives, and FP are false positives})$$

$$R = \frac{TP}{TP + FN} \quad (\text{where TP are true positives, and FN are false negatives})$$

- F-measure:

$$F = 2 \cdot \frac{P \cdot R}{P + R} \quad (\text{Harmonic mean of Precision and Recall})$$

- Mean Average Precision (MAP):

$$MAP = \frac{1}{n} \sum_{i=1}^n P_i$$

(To get the average precision for list of queries q_1, q_2, \dots, q_n .)

- Mean Average Precision (MAR):

$$MAR = \frac{1}{n} \sum_{i=1}^n R_i$$

(To get the average recall for list of queries q_1, q_2, \dots, q_n .)

- Mean Average Precision (MAF):

$$MAF = \frac{1}{n} \sum_{i=1}^n F_i$$

(To get the average F-measure for list of queries q_1, q_2, \dots, q_n .)

M 3.2 Archetype

Typesystem

For this milestone, you are required to use the typesystem `OAQATypes.xml` included in the archetype. Note however, you are only required to use the types defined in 3.2 and their supertype ancestry. You are of course welcome to extend the typesystem and/or use any other preexisting types. Most importantly, it is imperative that you use the `text` and `rank` attributes, as we will use them for evaluation.

For concepts, documents, triples, and snippets you must use type mappings described in M1 and M2. Also, you must use the full URI in the `uri` type.

Note, for yes/no answers we ignore the rank, and simply look for the yes/no answer in `text`. For factoid you will rank up to 5 answers, and for list you will rank up to 100.

BioASQ Type	UIMA Type
Answer	<ul style="list-style-type: none"> • <code>edu.cmu.lti.oaqa.type.answer.Answer</code> <ul style="list-style-type: none"> - <code>text</code> - <code>rank</code>

Figure 3.2: Typesystem Mapping

Web Services

You will use the same services as you did for M2 and M1.

Data

In the archetype you will also find `BioASQ-SampleData1B.json` containing an annotated version of the 29 sample questions shown in M0. We also provide you with a convenience class `JsonCollectionReaderHelper.java` to read from JSON format.

Bibliography

- [1] Georgios Balikas, Ioannis Partalas, Aris Kosmopoulos, Sergios Petridis, Prodromos Malakasiotis, Ioannis Pavlopoulos, Ion Androutsopoulos, Nicolas Baskiotis, Eric Gaussier, Thierry Artieres, and Patrick Gallinari. Evaluation framework specifications. Project deliverable D4.1, 05/2013 2013.