BioASQ

A challenge on large-scale biomedical semantic indexing and question answering

NCSR “D”

July 2016

Second Biocakt Group Meeting, Athens

Introduction

BioASQ: A challenge on large-scale biomedical semantic indexing and question answering

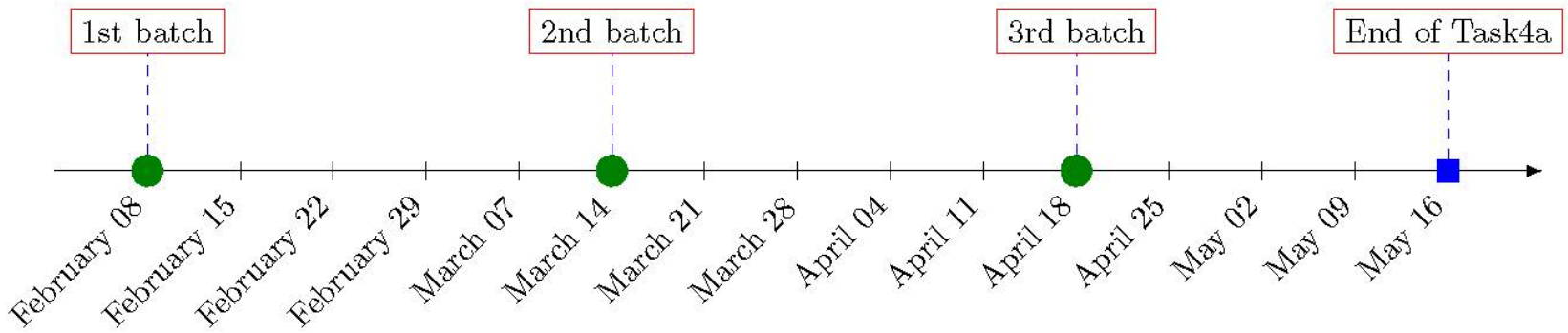
Systems competing on:

- Biomedical Semantic Indexing
  - Hierarchical Text Classification, Machine Learning
- Biomedical Information Retrieval
  - Document, Passage, RDF Triple Retrieval
- Biomedical Question Answering
  - Named Entity Recognition and Disambiguation
  - Information and Relation Extraction
  - Textual Entailment and Reasoning
  - Text Summarization, Natural Language Generation

Task A: Large-scale Biomedical Semantic Indexing

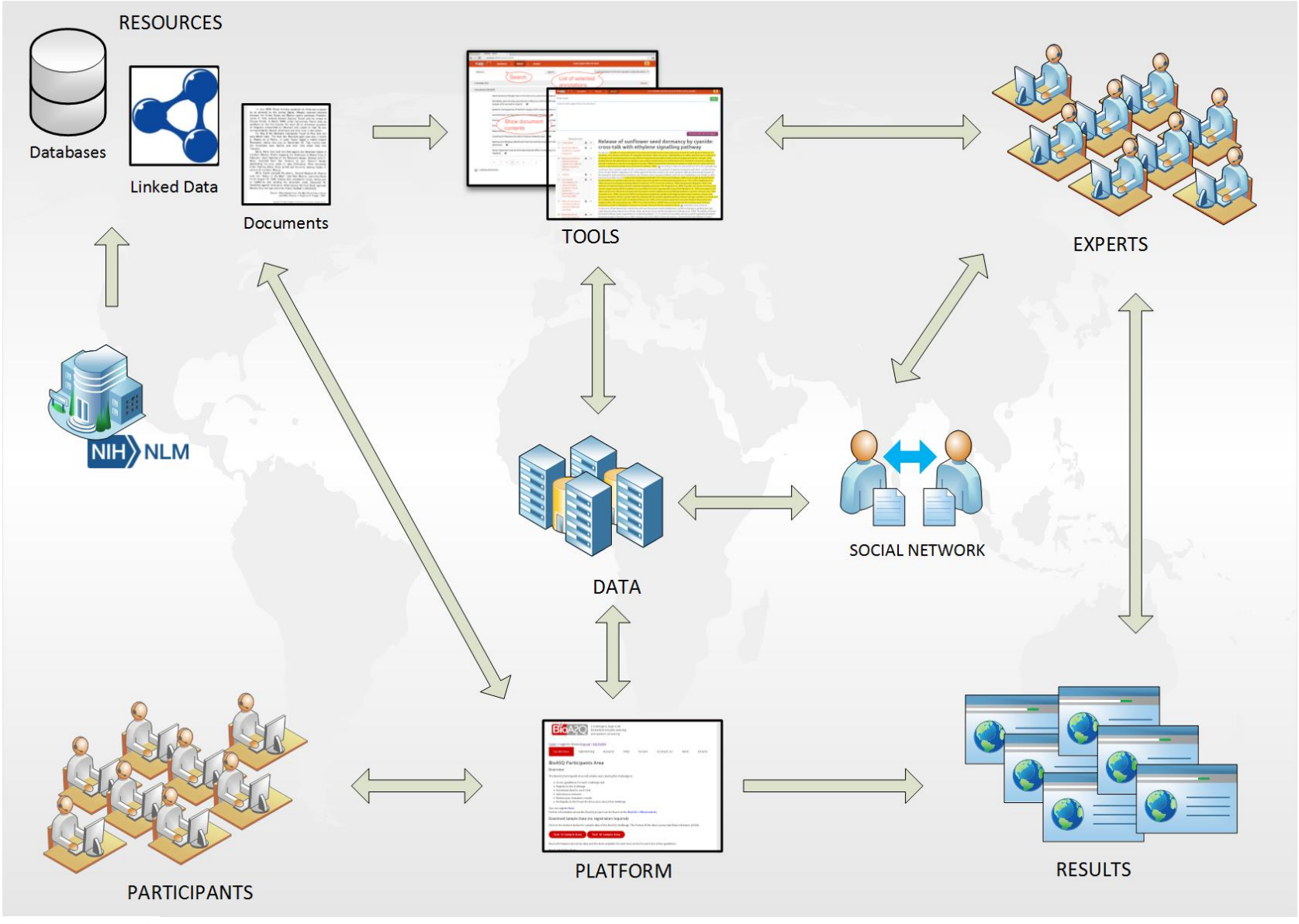
- Biomedical articles (PubMed)
- Medical Subject Headings (MeSH)

Task 4 A Schedule



Task A

BioASQ Architecture



Task A
Roll-out

Training data sets

All articles in PubMed with title, abstract and MESH

Tests

- Test data set creation
  - Articles in PubMed with title, abstract and NO MESH
  - MESH annotation in a few weeks (journal list)
- System results
  - MESH annotation in JSON format
  - limited time
- Results evaluation
  - MESH annotation available (incremental)
  - Evaluation measures

Task A
Evaluation Measures

Flat measures

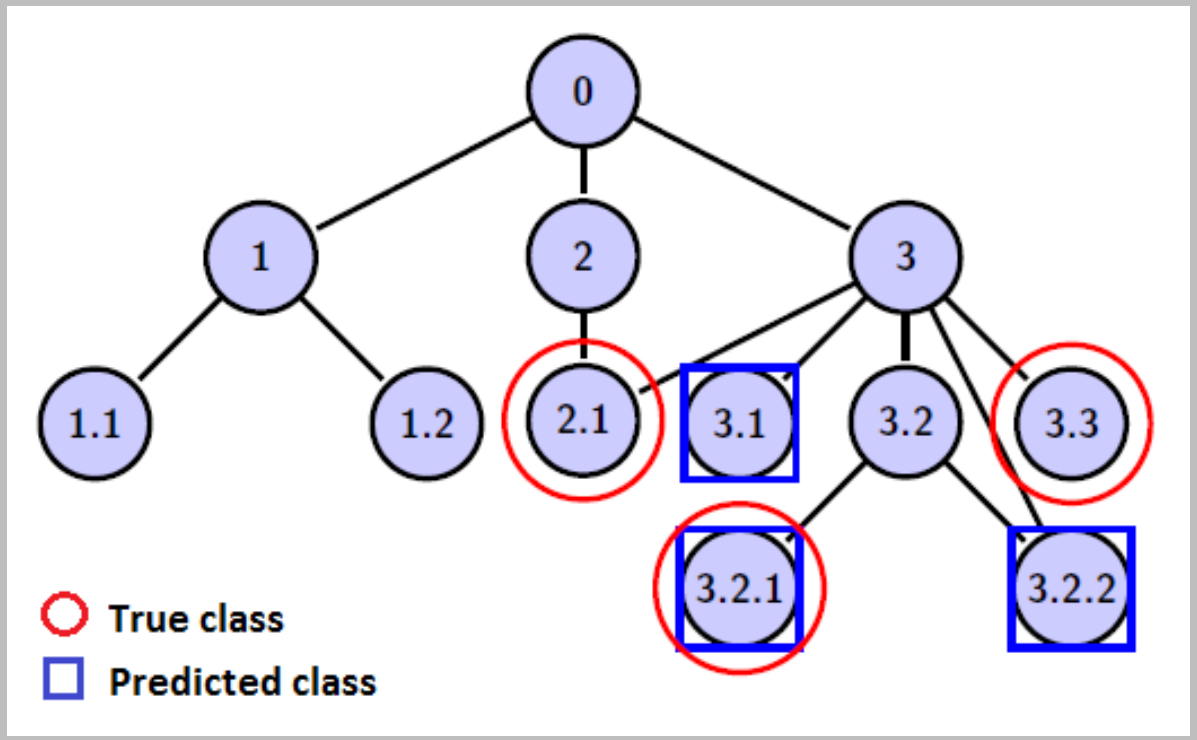
- Accuracy (Acc.)
- Example Based Precision/Recall/F-Measure (EBP, EBR, EBF)
- Macro Precision/Recall/F-Measure (MaP, MaR, MaF)
- Micro Precision/Recall/F-Measure (MiP, MiR, **MiF**)

Hierarchical measures

- Hierarchical Precision/Recall/F-Measure (HiP, HiR, HiF)
- Lowest Common Ancestor Precision/Recall/F-Measure (LCA-P, LCA-R, **LCA-F**)

Task A
Evaluation Measures

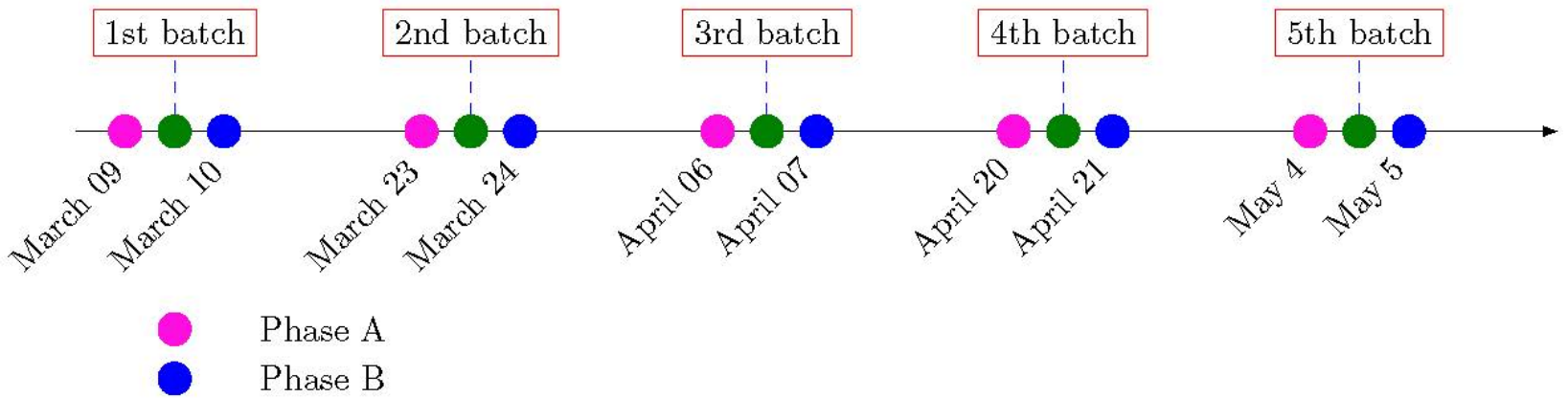
A simple Hierarchy



Task B : Biomedical Semantic Question Asnwering

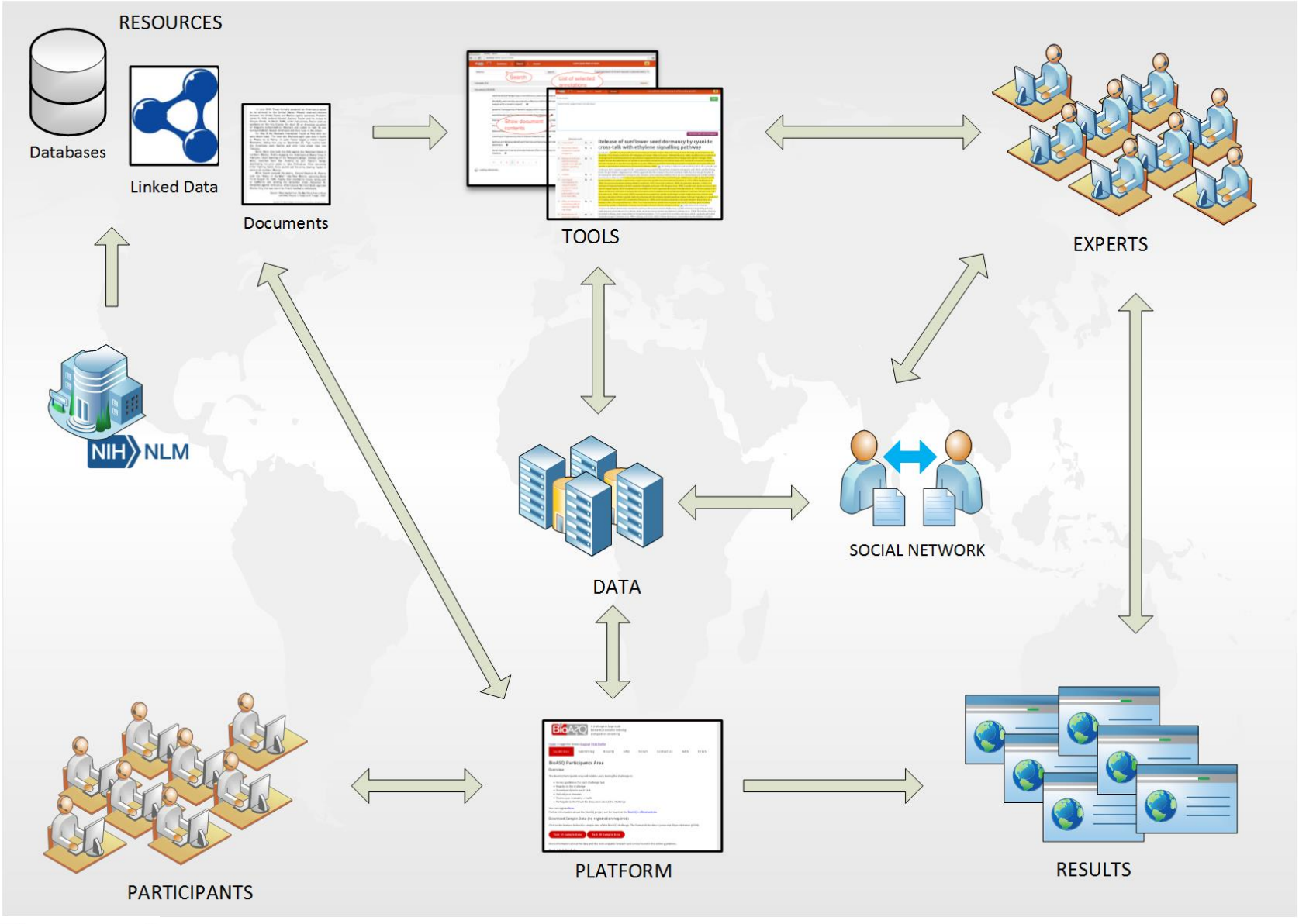
- Phase A : Semantic Indexing and Information Retrieval
  - Biomedical questions
  - Concepts, documents, snippets and RDF triples
- Phase B : Question Answering
  - Biomedical questions (annotated)
  - Exact and ideal answers

Task 4 B Schedule



Task B

BioASQ Architecture



Task B

Roll-out

Training data sets

1300 questions with enriched golden annotation and answers.

- Yes/No, Factoid, List, Summary

Tests

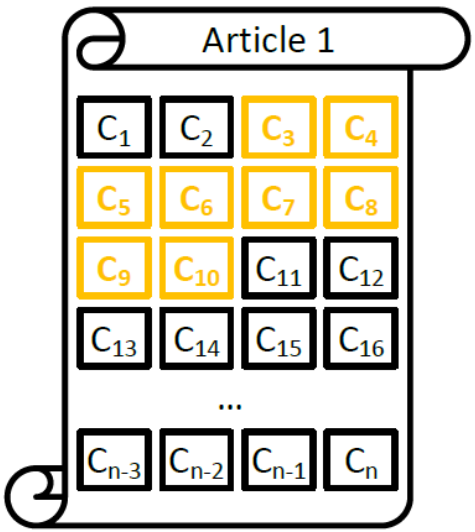
- Phase A (Search services provided)
  - Questions creation by biomedical experts Team (Annotation Tool)
  - System results : 10 documents, snippets, concepts, RDF triples
- Phase B
  - Questions and golden annotation
  - System results : exact and ideal answers
- Results evaluation
  - Enriched golden annotation and answers (Assessment Tool)
  - Evaluation measures

Task B
Evaluation Measures

Phase A

- Unordered : Mean Precision, Recall, F-Measure
- Ordered : Mean Average Precision (**MAP**), Geometric MAP

- Snippets case
  - article - offset pairs



Task B

Evaluation Measures

Phase B

Exact answers

- Yes/No : Accuracy (**Acc.**)
- Factoid : strict and lenient Acc, Mean Reciprocal Rank (**MRR**)
- List : Mean Precision, Recall, **F-measure**

Ideal answers : ROUGE, **Manual scores**

- Readability
- Recall
- Precision
- Repetition

Participation

Thank you !!!

References

- George Paliouras, Anastasia Krithara, Ioannis Kakadiaris, ”BioASQ Workshop Presentation, Toulouse, 9th September 2015”
- Tsatsaronis *et al*. ”An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition” BMC Bioinformatics (2015)
- http://www.bioasq.org/(2016)
- Kosmopoulos A, Partalas I, Gaussier E, Paliouras G, Androutsopoulos I. Evaluation measures for hierarchical classification: a unified view and novel approaches. Data Mining and Knowledge Discovery. 2014;29:146.
- Balikas *et al*. ”D4.1 Evaluation Framework Specifications” (2013)
- Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas ”Mining Multi-label Data” Data Mining and Knowledge Discovery Handbook, Springer (2010)

Appendix
Evaluation Measures Task A

Flat measures

- Micro Precision/Recall/F-Measure (MiP, MiR, **MiF**)

$$MiF1 = \frac{2 * MiP * MiR}{MiP + MiR},$$

$$MiP = \frac{\sum_{i=1}^{|C|} tp_{c_i}}{\sum_{i=1}^{|C|}(tp_{c_i} + fp_{c_i})}, MiR = \frac{\sum_{i=1}^{|C|} tp_{c_i}}{\sum_{i=1}^{|C|}(tp_{c_i} + fn_{c_i})}$$

tp_{c_i}, fp_{c_i} and fn_{c_i} are true positives, false positives and false negatives for class c_i of class set C.

Appendix
Evaluation Measures Task A

Hierarchical measures

- Lowest Common Ancestor Precision/Recall/F-Measure (LCA-P, LCA-R, **LCA-F**)

$$LCaF = \frac{2 * LCaP * LCaR}{LCaP + LCaR}$$

$$LCaP = \frac{|\hat{Y}_{aug} \cap Y_{aug}|}{|\hat{Y}_{aug}|}, LCaR = \frac{|\hat{Y}_{aug} \cap Y_{aug}|}{|Y_{aug}|}$$

Y_{aug} and $\hat{Y}_{aug}$ are augmented sets of the true and the predicted classes respectively

Appendix
Evaluation Measures Task B

Phase A

- Mean Average Precision (**MAP**)

$$MAP = \frac{1}{n} \cdot \sum_{i=1}^{n} AP_i,$$

$$AP = \frac{\sum_{r=1}^{|L|} P(r) \cdot rel(r)}{|L_R|}$$

n the number of questions, L the list of items, L_R the set of relevant items, $P(r)$ precision when keeping only first r items, and $rel(r)$ equals 1 if the r-th item is relevant and 0 otherwise.

Appendix
Evaluation Measures Task B

Phase B

Exact answers

- Yes/No : Accuracy **Acc** $= \frac{c}{n}$
  where n the number of questions.
- Factoid : Mean Reciprocal Rank **MRR**: $= \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{1}{r(i)}$
  where $r(i) = j$ if j-th the topmost position of golden answer, $r(i) \rightarrow +\infty$ otherwise.
- List : Mean **F-measure** = $\frac{1}{n} \cdot \sum_{i=1}^{n} F(i)$