CLEF 2014 – Conference and Labs of the Evaluation Forum

BioASQ workshop

HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

HPI in-memory-based database system in Task 2b of BioASQ

Mariana Neves

September 16th, 2014

Outline

- Overview of participation
- Architecture of the the system
- Results
- Discussion and future work

Outline

- Overview of participation
- Architecture of the system
- Results
- Discussion and future work

Participation

- Phase A of task 2b
  - Given: list of 100 questions and respective type
  - Required:
    - a list of relevant concepts:GO, DO, MeSH, Jochem and Uniprot
    - a list of relevant documents from PubMed (PMID)
    - a list of relevant snippets: PMID, start and end sections and offsets in the documents, and the text of the snippet

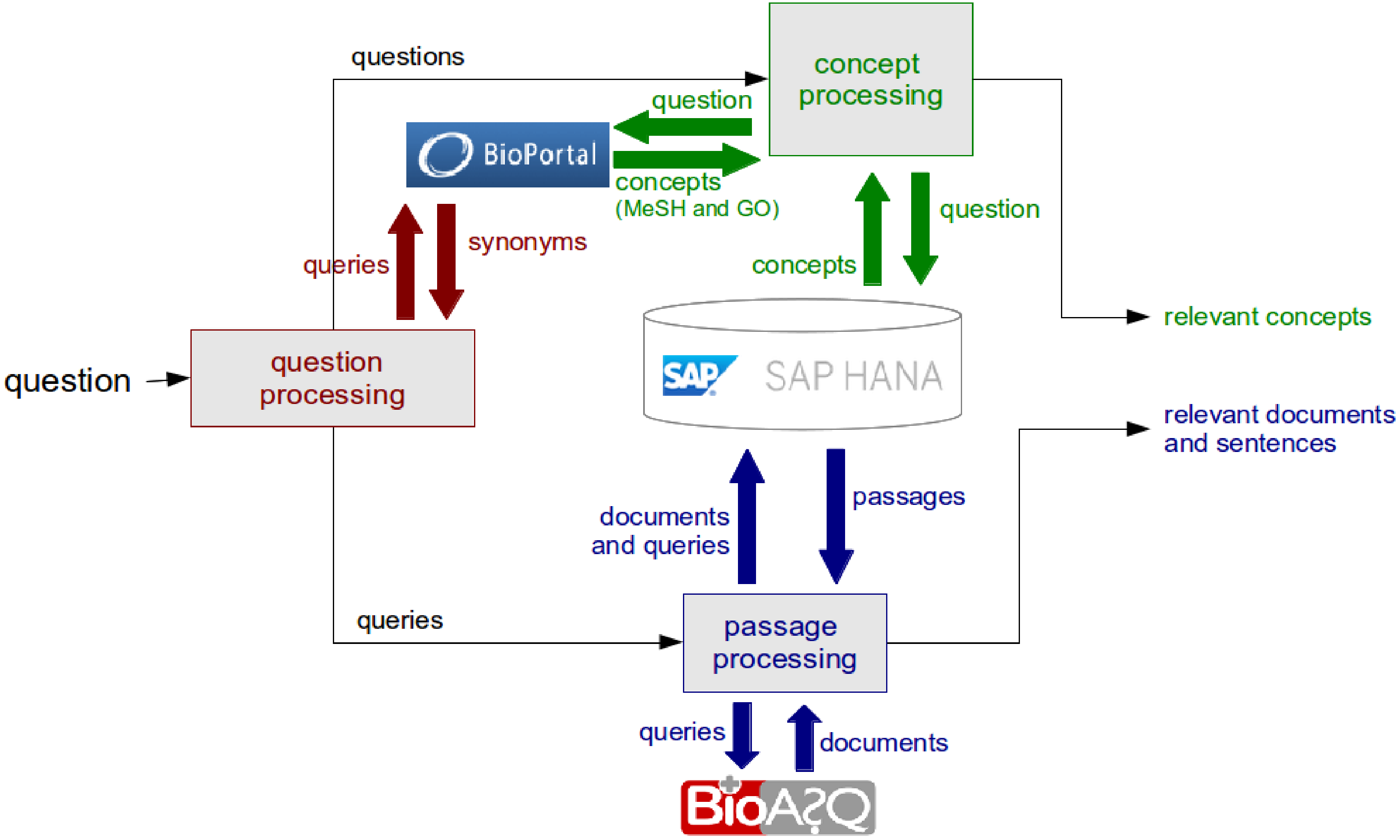
Participation

- Phase A of task 2b
  - 5 batches of 100 questions each:
    - Participation in batches 2,3 and 4
    - Concepts only in batches 3 and 4

Outline

- Overview of participation
- Architecture of the system
- Results
- Discussion and future work

Architecture of the system



Question processing

- Stanford CoreNLP: sentence splitting, tokenization, POS tagging, chunking
- Query construction
  - Based on tokens and chunks
  - Removal of numerals (POS „CD‟), stopwords, length less than 3
  - „Is Rheumatoid Arthritis more common in men or women?‟
    - Rheumatoid, Arthritis, more, common, men, women
    - Rheumatoid Arthritis, more, common, men or women

Question processing

- Query expansion
  - BioPortal
  - All returned synonyms and defintion up to length 20
  - Ignored synonyms with symbols/punctuations
    - „Homo sapiens (living organism) [Ambiguous]"
  - Weights
    - 0.5 for terms not matching synonyms
    - Otherwise:

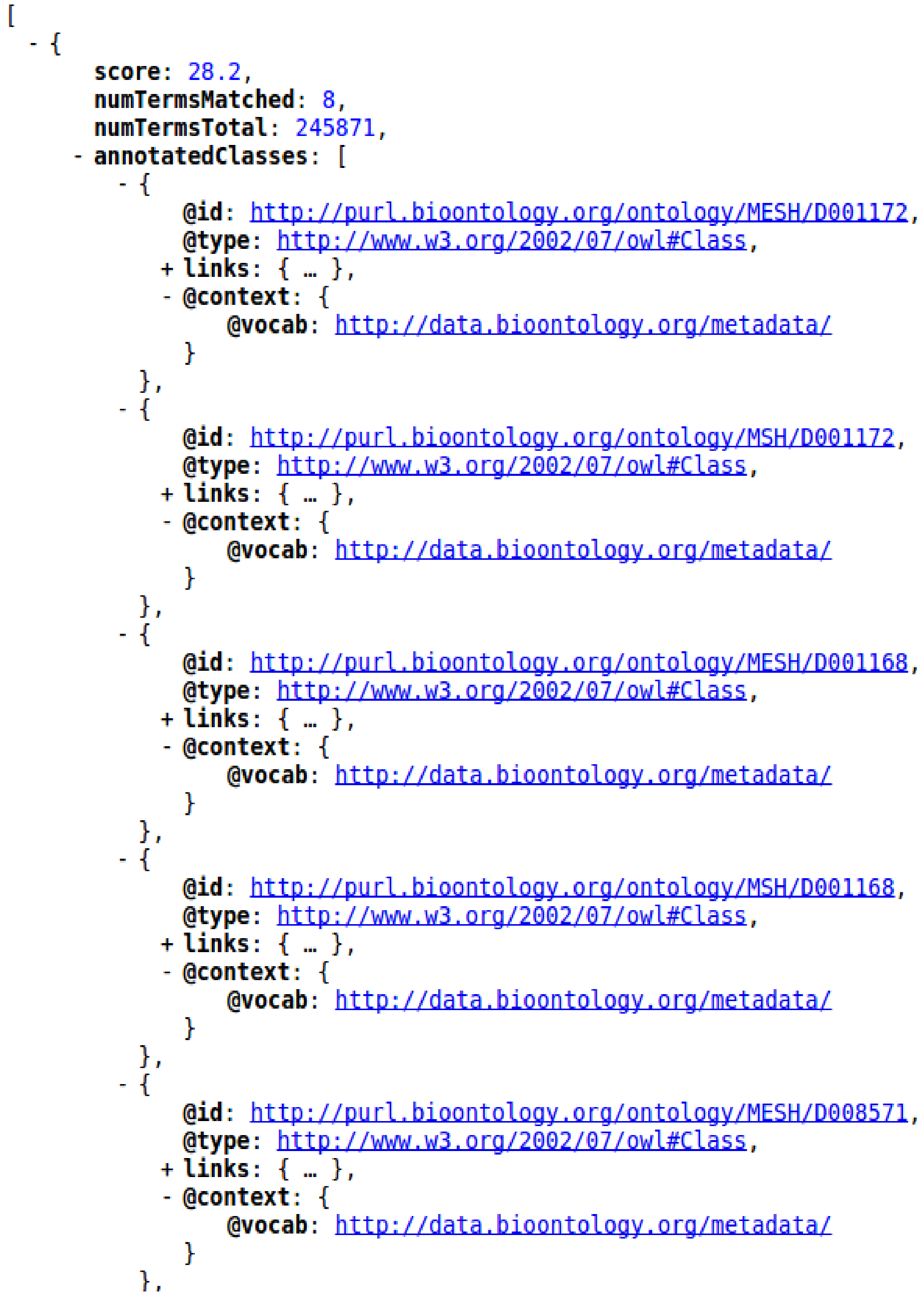
$$weight = 1 - \frac{\#MatchesToken}{\#MatchesTotal}$$

Concept processing

- BioPortal Recommender (MeSH and GO)
  - Queries based on full text of the question
- SAP HANA database:
  - Compiled dictionaries
    - Jochem: lines “ID” (identifiers) and “TM” (terms);
    - DO, MeSH and GO (OBO files): fields “id” (identifiers), “name” (names) and “synonym” (synonyms);
    - SwissProt: lines “ID” (identifiers), “DE” (description) and “GN” (gene names)
  - Ignore terms whose length less than 3, stopwords and Greek letters

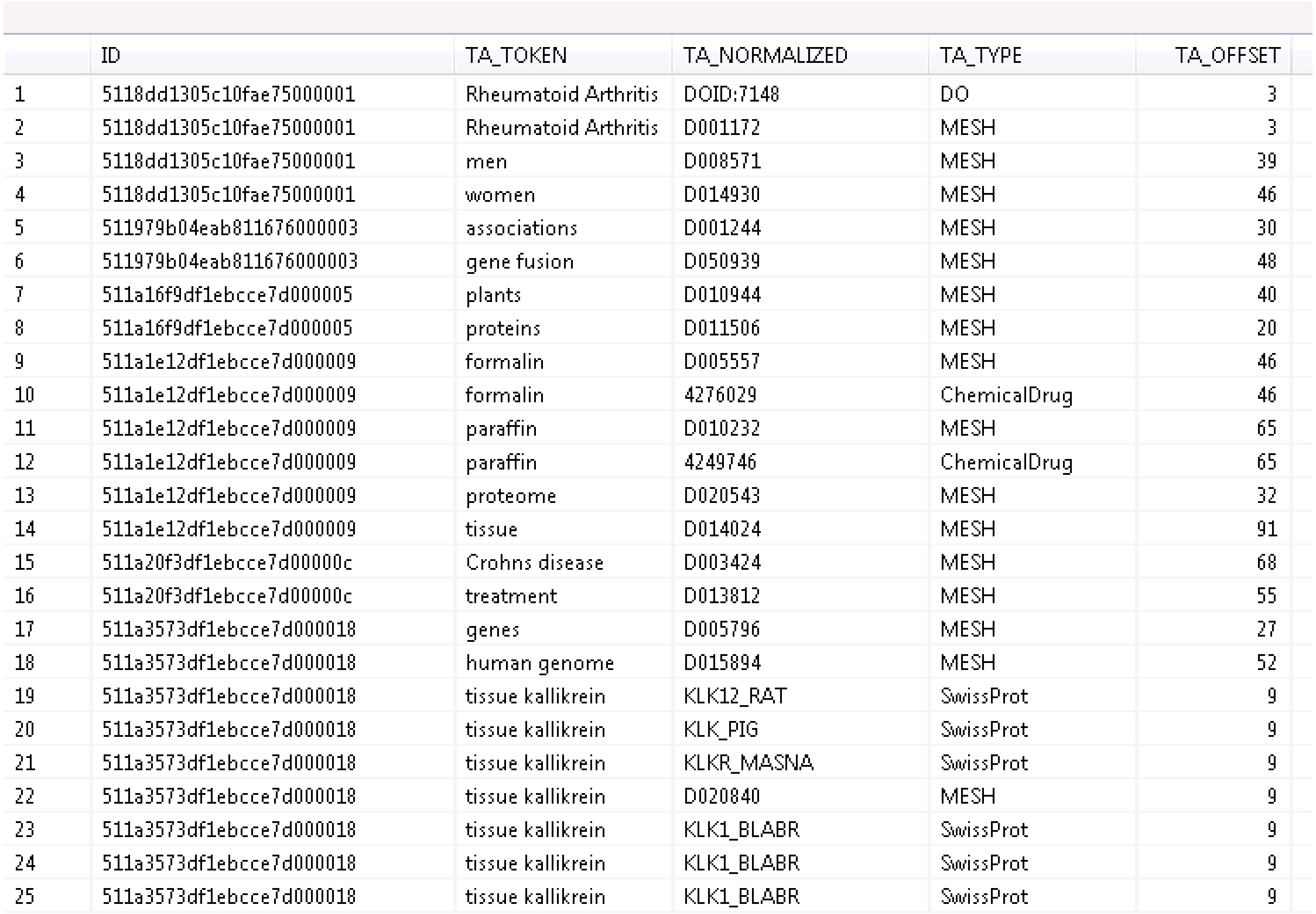
Concept processing

BioPortal



```
[
  - {
        score: 28.2,
        numTermsMatched: 8,
        numTermsTotal: 245871,
      - annotatedClasses: [
          - {
                @id: http://purl.bioontology.org/ontology/MESH/D001172,
                @type: http://www.w3.org/2002/07/owl#Class,
              + links: { … },
              - @context: {
                    @vocab: http://data.bioontology.org/metadata/
                }
            },
          - {
                @id: http://purl.bioontology.org/ontology/MSH/D001172,
                @type: http://www.w3.org/2002/07/owl#Class,
              + links: { … },
              - @context: {
                    @vocab: http://data.bioontology.org/metadata/
                }
            },
          - {
                @id: http://purl.bioontology.org/ontology/MESH/D001168,
                @type: http://www.w3.org/2002/07/owl#Class,
              + links: { … },
              - @context: {
                    @vocab: http://data.bioontology.org/metadata/
                }
            },
          - {
                @id: http://purl.bioontology.org/ontology/MSH/D001168,
                @type: http://www.w3.org/2002/07/owl#Class,
              + links: { … },
              - @context: {
                    @vocab: http://data.bioontology.org/metadata/
                }
            },
          - {
                @id: http://purl.bioontology.org/ontology/MESH/D008571,
                @type: http://www.w3.org/2002/07/owl#Class,
              + links: { … },
              - @context: {
                    @vocab: http://data.bioontology.org/metadata/
                }
            },
```

Concept processing



	ID	TA_TOKEN	TA_NORMALIZED	TA_TYPE	TA_OFFSET
1	5118dd1305c10fae75000001	Rheumatoid Arthritis	DOID:7148	DO	3
2	5118dd1305c10fae75000001	Rheumatoid Arthritis	D001172	MESH	3
3	5118dd1305c10fae75000001	men	D008571	MESH	39
4	5118dd1305c10fae75000001	women	D014930	MESH	46
5	511979b04eab811676000003	associations	D001244	MESH	30
6	511979b04eab811676000003	gene fusion	D050939	MESH	48
7	511a16f9df1ebcce7d000005	plants	D010944	MESH	40
8	511a16f9df1ebcce7d000005	proteins	D011506	MESH	20
9	511a1e12df1ebcce7d000009	formalin	D005557	MESH	46
10	511a1e12df1ebcce7d000009	formalin	4276029	ChemicalDrug	46
11	511a1e12df1ebcce7d000009	paraffin	D010232	MESH	65
12	511a1e12df1ebcce7d000009	paraffin	4249746	ChemicalDrug	65
13	511a1e12df1ebcce7d000009	proteome	D020543	MESH	32
14	511a1e12df1ebcce7d000009	tissue	D014024	MESH	91
15	511a20f3df1ebcce7d00000c	Crohns disease	D003424	MESH	68
16	511a20f3df1ebcce7d00000c	treatment	D013812	MESH	55
17	511a3573df1ebcce7d000018	genes	D005796	MESH	27
18	511a3573df1ebcce7d000018	human genome	D015894	MESH	52
19	511a3573df1ebcce7d000018	tissue kallikrein	KLK12_RAT	SwissProt	9
20	511a3573df1ebcce7d000018	tissue kallikrein	KLK_PIG	SwissProt	9
21	511a3573df1ebcce7d000018	tissue kallikrein	KLKR_MASNA	SwissProt	9
22	511a3573df1ebcce7d000018	tissue kallikrein	D020840	MESH	9
23	511a3573df1ebcce7d000018	tissue kallikrein	KLK1_BLABR	SwissProt	9
24	511a3573df1ebcce7d000018	tissue kallikrein	KLK1_BLABR	SwissProt	9
25	511a3573df1ebcce7d000018	tissue kallikrein	KLK1_BLABR	SwissProt	9

Passage processing

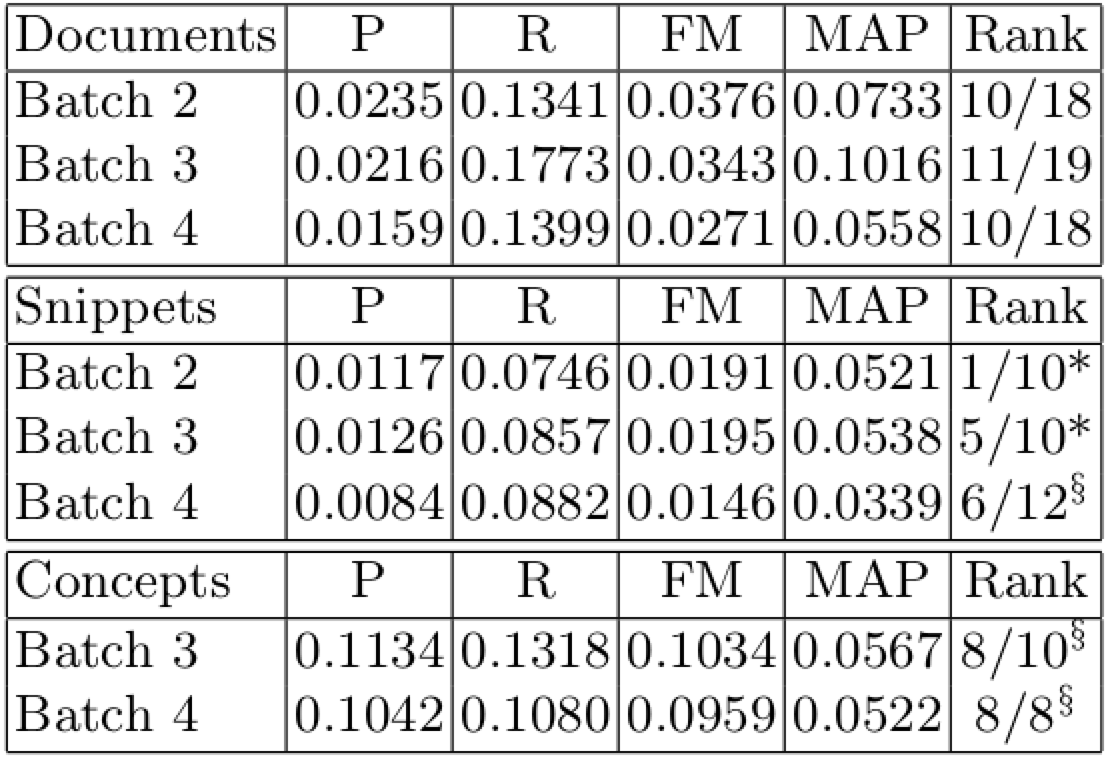
- Document Retrieval: BioASQ PubMed service
  - W/ and w/o query expansion; OR and AND operators
  - 500 top ranked documents
  - Only titles and abstracts (90% of passages on training data)
- Document indexing (SAP HANA):
  - Sentence splitting and tokenization
- Passage retrieval:
  - Token-based fuzzy search (>90% similarity)
  - Sentence ranked by score: string similarity + weights
  - Only documents returned by BioASQ for the question
  - Top 100 ranked passages

Outline

- Overview of participation
- Architecture of the the system
- Results
- Discussion and future work

Results

- Test data (batches 2, 3, 4)



Documents	P	R	FM	MAP	Rank
Batch 2	0.0235	0.1341	0.0376	0.0733	10/18
Batch 3	0.0216	0.1773	0.0343	0.1016	11/19
Batch 4	0.0159	0.1399	0.0271	0.0558	10/18

Snippets	P	R	FM	MAP	Rank
Batch 2	0.0117	0.0746	0.0191	0.0521	1/10*
Batch 3	0.0126	0.0857	0.0195	0.0538	5/10*
Batch 4	0.0084	0.0882	0.0146	0.0339	6/12§

Concepts	P	R	FM	MAP	Rank
Batch 3	0.1134	0.1318	0.1034	0.0567	8/10§
Batch 4	0.1042	0.1080	0.0959	0.0522	8/8§

* ranked higher than the Top 100 and Top 50 baselines

§ no system outperformed none of the two baselines

Outline

- Overview of participation
- Architecture of the the system
- Results
- Discussion and future work

Discussion and error analysis

- Higher recall than precision
  - 100 top concepts, documents, snippets
  - Definition of a threshold

- Concept retrieval
  - Ex. “Is Rheumatoid Arthritis more common in men or women?” (5118dd1305c10fae75000001)
    - FPs: “D001172” (Arthritis, Rheumatoid), “D014930” (Women)
    - FNs: “D001171” (Arthritis, Juvenile Rheumatoid) and “D015535” (Arthritis, Psoriatic)

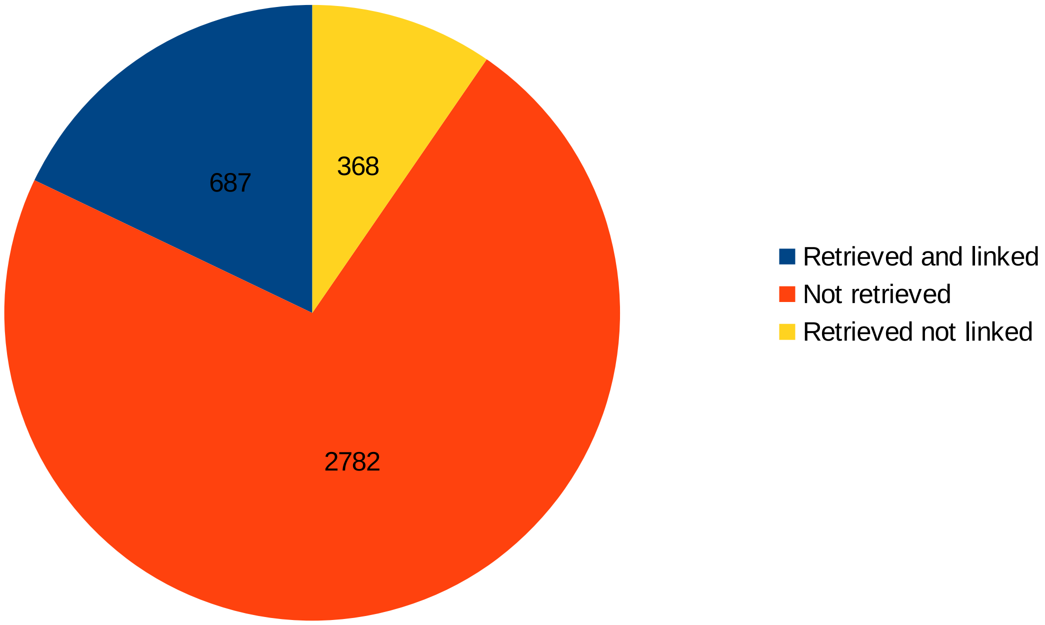
Discussion and error analysis

- Concept retrieval
  - "What is the effect of TRH on myocardial contractility?" (5160193d298dcd4e51000039)
    - PH4H_DROME ("TRH", *D. melanogaster*)
  - "Describe the known functions for the prothymosin alpha c-terminal peptide?" (51be03c4047fa84d1d000004)
    - FN: PTMA_HUMAN ("Prothymosin alpha")
    - FP: PAHO_MOUSE ("C-terminal peptide")
  - "Are there any DNMT3 proteins present in plants?" (511a16f9df1ebcce7d000005)
    - CMT1_ARATH, CMT2_ARATH, CMT3_ARATH
    - „DNA (cytosine-5)-methyltransferase CMT3"

Discussion and error analysis

- Document and passage retrieval
  - Dependent on our queries and BioASQ services

False positives:



Future works

- Additional search engines
  - GeneView (Humboldt-Universität Berlin)
  - PubMed
- Question processing
  - Semantic role labeling
  - Expected answer extraction
  - Question taxonomy
- Answer processing
  - Exact answer
  - Summarization

Thank you!

Questions?

- Acknowledgements:
  - HPI Research School

- Contact:
  - marianalaraneves@gmail.com