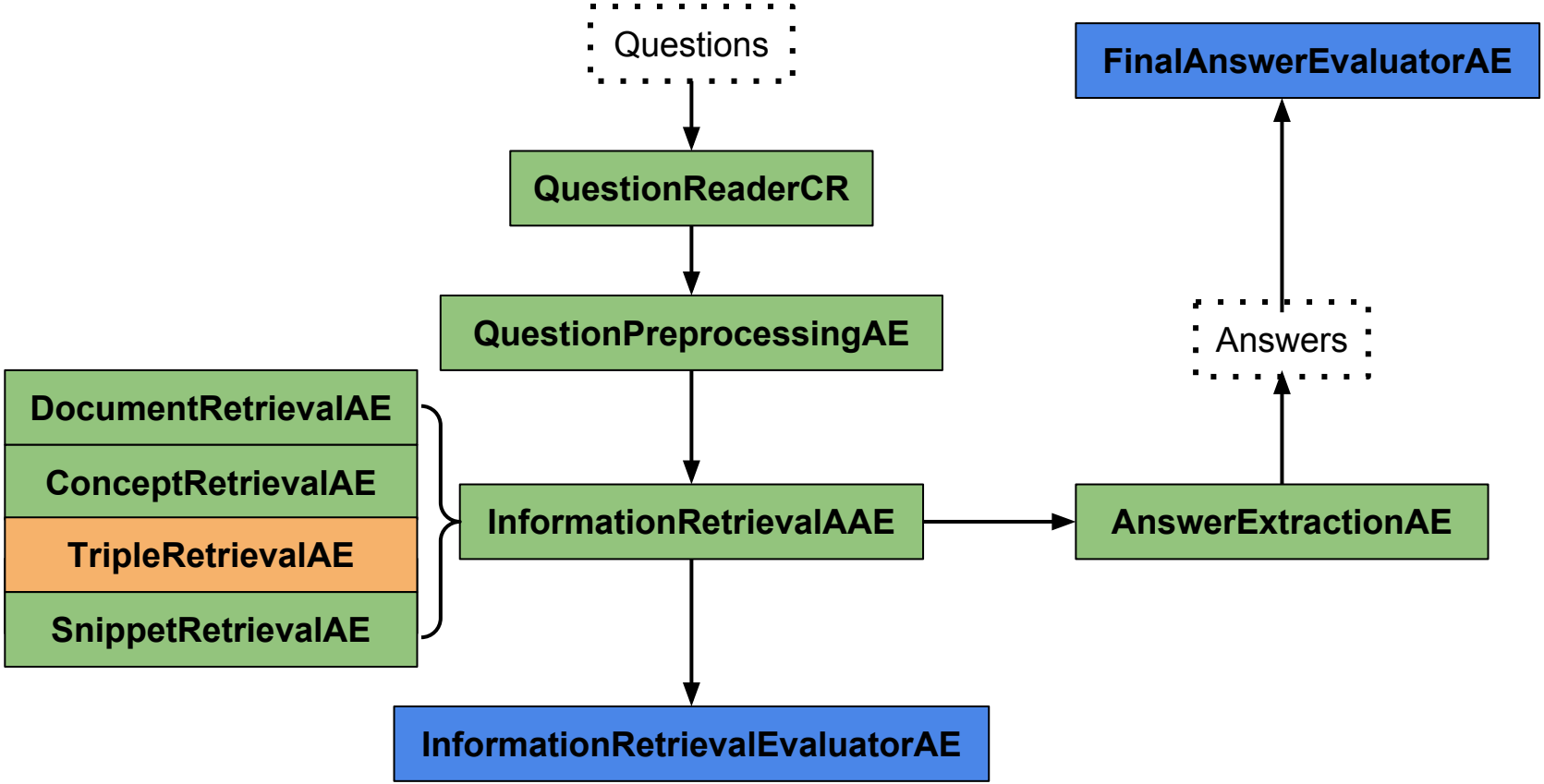
Cooper Spaghetti

11791 - TEAM 03

Junjia He, Chu-Cheng Lin,
Mohammad Gowayyed, Han Zhang

Overview


Questions
QuestionReaderCR
QuestionPreprocessingAE
DocumentRetrievalAE
ConceptRetrievalAE
TripleRetrievalAE
SnippetRetrievalAE
InformationRetrievalAAE
InformationRetrievalEvaluatorAE
AnswerExtractionAE
Answers
FinalAnswerEvaluatorAE

Results Highlights

	Precision	Recall	F1	MAP	GMAP
Concept	0.1548	0.3824	0.2203	0.3589	0.1761
Document	0.0130	0.0923	0.0228	0.0715	0.0260
Snippet	0.0015	0.0083	0.0026	0.0005	0.0104
Match	0.0091	0.2800	0.0176	0.0158	0.0170

More Data

- We only have **8** list questions in *BioASQ-SampleData1B.json*

- We extracted **91** list questions out of *BioASQ-trainingDataset2b.json* to use in experimentation.

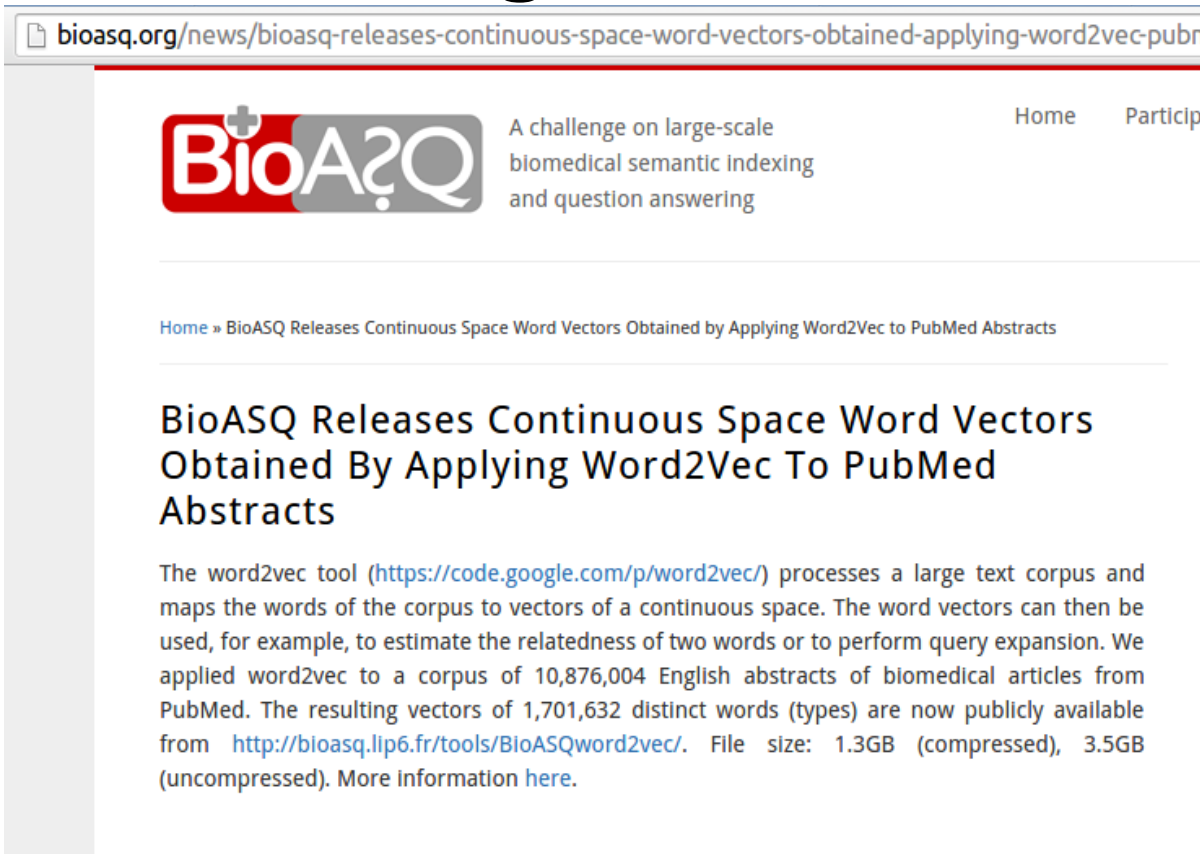
NLP Resources

- Stanford Stemmer
- Stanford Pos tagger
- Abner NER
- lingpipe chunker
- word2vec

Word Embeddings

1. real-valued dense representation of tokens
2. captures semantic regularities in a language
    a. similar words have similar embeddings
3. word2vec
    a. fast training
    b. scales well

Word Embeddings



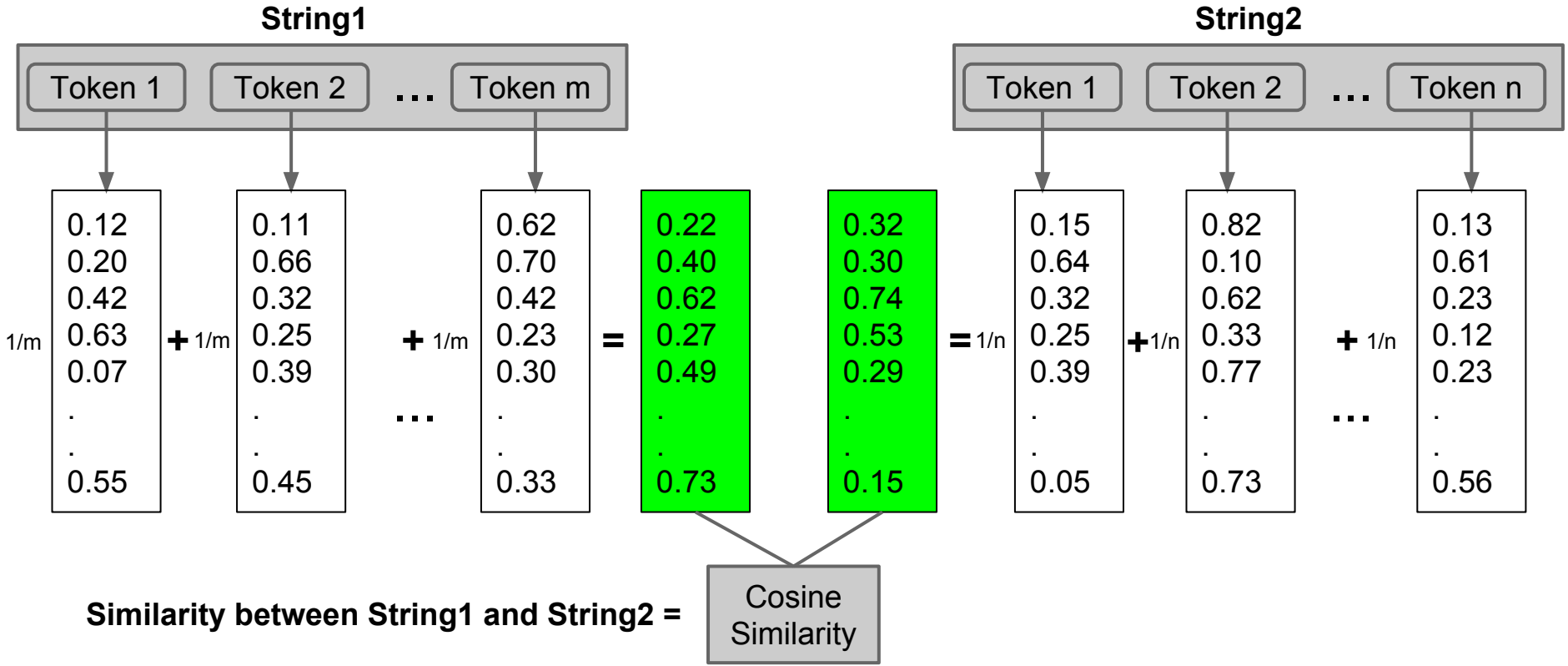
pubvec online service

- We implemented a web service to speed up working with word vectors
- Ruby on Rails
- 1,701,632 words (types)
- We included only words that start with either a letter or a digit: 1,456,789 words
- MySql database with index on words

pubvec online service

- Given a word:
  - return its 200-length vector
- Given a string and n
  - return the closest n words (in cosine similarity)
  - used for query expansion
- Given two strings
  - returns the cosine similarity between their two vectors
  - a vector of a string is the addition of its words vectors
  - used in snippet retrieval and answer extraction

Computing similarity on the server



Query Expansion - Local

1. <original_query> OR <*NE AND NE ...*> OR <*concept AND concept ...*> OR <*bigrams*>
   - precision 0.1553 -> 0.1448, recall stays the same
2. <noun_unigram AND syn…> OR <noun_bigram AND syn…>
   - Stanford POS tagging
   - Use UMLS Metathesaurus to retrieve synonyms (atoms)
   - precision 0.1553 -> 0.0483, **recall 0.1070 -> 0.1973 :)**
3. <noun_unigram OR noun_unigram OR noun_bigram...>
   - A fallback strategy for query formulation

Query Expansion - Local (cont.)

- Fallback
  - First, use expansion scheme 2 (UMLS and synonyms)
  - If retrieved documents are empty *(it actually happens!)*, use expansion scheme 3 (which guarantees to find something, in our case)

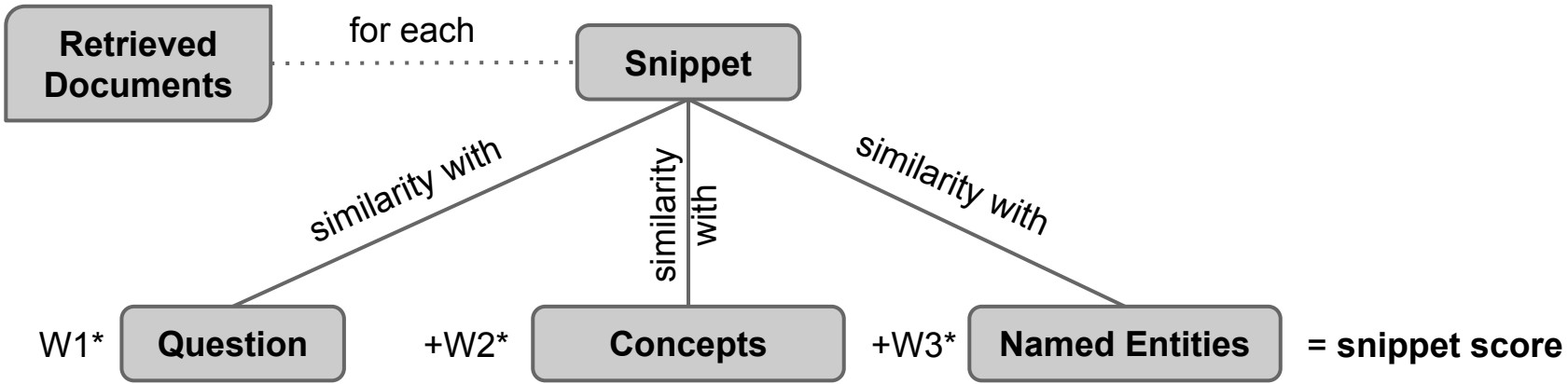
- *Recall is everything*, we don't care the precision or ranking. (All snippets/sentences are consdered the same after document retrieval)

Concept Retrieval

- replace “2014” with “2012”

- use threshold of 0.15

Snippet Retrieval



*Distances are calculated using our *pubvec* web service.

Snippet Retrieval

- For each section in a retrieved document:
  - For each snippet
    - Calculate score as described using the *pubvec* web service.
- Sort all the snippets and select the *TOP_K* ones.
- Experimented with *TOK_K* = 10, 50, and 100.
- 100 always gives better results.

Snippet Retrieval

Results

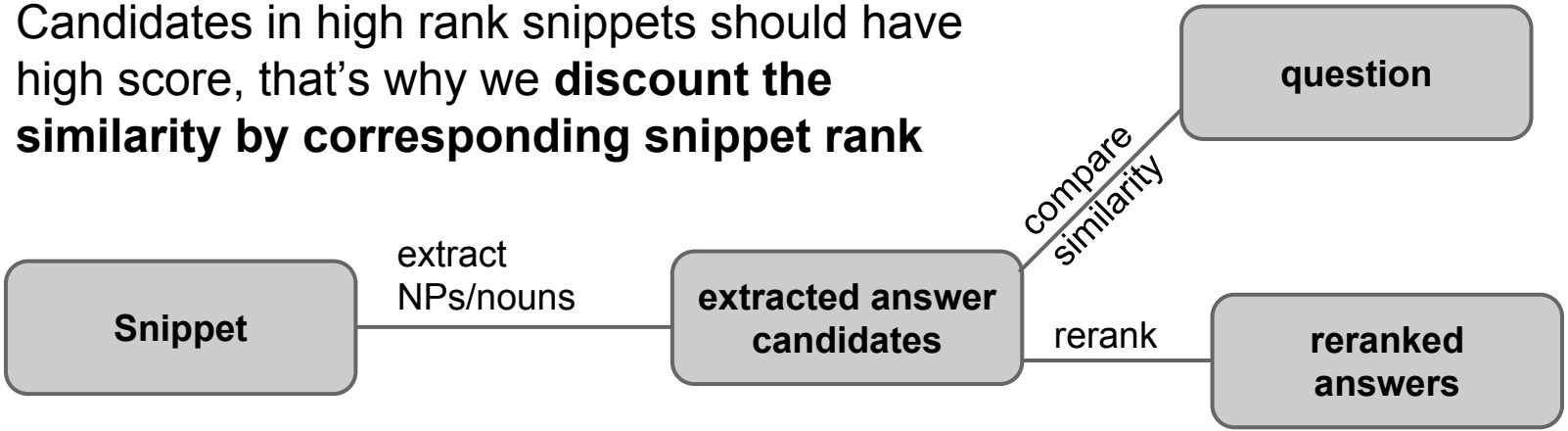
Only set one weight to one while keeping others zero

Snippet Retrieval	Precision	Recall	F1	MAP	GMAP
QuestionWeight=1	0.0364	0.0049	0.0087	0.0022	0.0110
ConceptsWeight=1	0.0364	0.0049	0.0087	0.0026	0.0110
EntityWeight=1	0.0182	0.0025	0.0043	0.0034	0.0109

However, later you'll see...

Answer Extraction

- NP extraction: consecutive NN*s (up to 3)
- answers that appear in the question (exactly match) are not considered
- Cosine similarity in the representation space
- Candidates in high rank snippets should have high score, that's why we **discount the similarity by corresponding snippet rank**



Discounted Similarity Score

- Inspired by Discounted Cumulative Gain

$$Score = \frac{2^{sim} - 1}{\log_2 (rank + 1)}$$

- Score varies from [0, 1] - *which is good!*

Evaluation

1. intermediate results --- exact match
2. final answer --- soft match

   counts if gold answer is a substring of the retrieved answer
3. Example

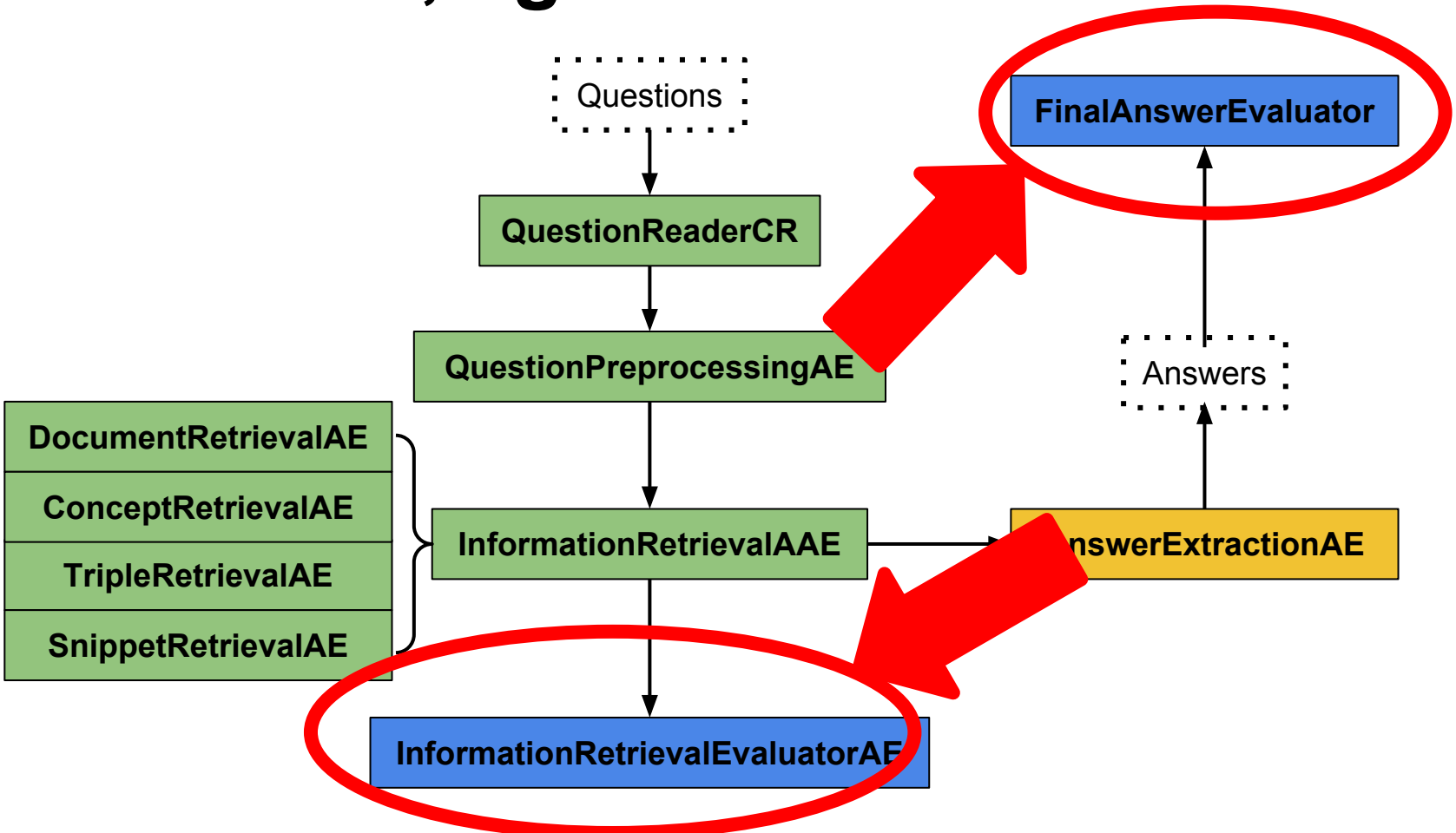
   golden: *AATAAA*

   retrieved: *sequence AATAAA, consensus sequence AATAAA*

   strength: can capture plural nouns

   weakness: say if there is another gene AAATAAA

Overview, again



Evaluation

1. Intermediate & final answer evaluations very similar
   a. load ground truth in initialize()
   b. compare retrieved list of results to ground truth
   c. print scores
2. actual computation of metrics done in class Stats

Refactoring IR Evaluator: Before

- Computation of metrics scattered in InformationRetrievalEvaluator
- Business logic intertwined with UIMA workflow
- Duplicate code for Document/Concept/Triple

Refactoring IR Evaluator: After

- New helper class Stats provides a layer of abstraction
  - computes TP/FP/FN/AP given golden/predicted pairs
- Calculation of recall/precision/… consolidated in a class
- Separation of UIMA logic and domain logic
- Effortlessly extended for Snippet evaluation

Error Analysis

Query: what prominent sequence consensus polyadenylation site

golden:

AATAAA

AAUAAA

INTERSECTION OF answers IS: 2.0

Answers retrieved in our early development

polyadenylation consensus sites 0.8956057328374761

consensus polyadenylation sites 0.8956057328374761

consensus sequence **AATAAA** 0.8657756389900734

...

AATAAA: rank 71

consensus AAUAAA: rank 12

Error Analysis (cont.)

1. Answers are segments of the question
    a. because they have high cosine similarity with the question in the vector space
    b. possible solution: explicitly prohibit the answers to contain words in the question (work in progress)
2. Long answers hurt performance

Error Analysis

Query: what prominent sequence
consensus polyadenylation site

golden:

AATAAA

AAUAAA

AATAAA: rank 71 -> rank 6

AAUAAA: rank 100+ -> rank 18

```
RUNNING DOCUMENT RETRIEVAL
Retrieved Document: 100
RUNNING SNIPPET RETRIEVAL
RUNNING ANSWER EXTRACTION
Loading default NLPBA tagging module...
    1: sequence ATTAAA -    0.450865
    2: ATTAAA - 0.396133
    3: sequence AATAAA -    0.359551
    4: polyadenylation motif -  0.327834
    5: polyadenylation signal - 0.326405
    6: AATAAA - 0.314711
    7: consensus AAUAAA -   0.300240
    8: AAUAAA polyadenylation - 0.299940
    9: AAUAAA hexanucleotide -  0.251910
    10: polyadenylation sites - 0.241989
    11: motif - 0.235815
    12: sequence AAUAAA -   0.227017
    13: polyadenylation region -    0.212556
    14: consensus hexanucleotide -  0.209286
    15: splice site -   0.193688
    16: hexanucleotide sequence -   0.193295
    17: splicing enhancer - 0.181991
    18: AAUAAA -    0.177006
```

Project Management

- All management is done using quip.com
  - SCRUM backlog
  - Sprint planning
  - Messaging between team members
- We don’t use github issues
- Peer programming

Division of Labor

- Mohammad Gowayyed
  - word2vec web service, concept and snippet retrieval, experiments
- Junjia He
  - query expansion, document retrieval, experiments
- Chu-Cheng Lin
  - answer extraction, evaluation code, experiments
- Han Zhang
  - UMLS web service, triple retrieval, experiments

Future Work

- Optimize the query expansion on the server side
- Support other types of questions
- Better representation of snippets using word embeddings
- Better ranking schemes

Thanks