Learning to Answer Biomedical Factoid & List Questions: OAQA at BioASQ 3B

Zi Yang, Niloy Gupta, Xiangyu Sun, Di Xu, Chi Zhang, Eric Nyberg

Language Technologies Institute
School of Computer Science
Carnegie Mellon University

{ziy, ehn}@cs.cmu.edu
Overview

• Past history
  – First-time BioASQ participation
  – Built system for TREC Genomics QA & CLEF QA4MRE

• Hypothesis and preparation
  – Learning from past dryrun/1B/2B development set
  – Careful design of a flexible and extensible architecture, coupled with continuous, incremental experimentation and optimization over various combinations of existing state-of-the-art components
  – April 2 to Jun 10 (targeting for batches 3-5), 717 experiments have been recorded by the experiment database
    • Testing: 422, training: 167, caching: 80
Architecture

Components

Bio Components

BioQA – Biomedical domain specific modules

BaseQA – QA type system, domain independent QA modules, evaluators, etc.

ECD/CSE – Extension to declarative descriptors of components and pipeline, component composition, experiment logging, etc.

TREC Genomics QA

CLEF QA4MRE

CLEF BioASQ QA – Task specific modules (e.g. GoPubMed services)

Enterprise QA for drug discovery

Course project system

UIMA

JAVA
Factoid & List Question Answering for Phase B

Input question
- Question parsing
  - Question concept identification
  - Lexical answer type extraction
  - Answer type prediction

Relevant snippets
- Snippet parsing
  - Snippet concept identification
  - Concept retrieval
  - Concept merging

Candidate answer variant generation

Candidate answer variant merging

Answer scoring and ranking

Answer pruning

Answer
Question and Answer Type Prediction

• Answer type definition
  – UMLS semantic types + QUANTITY + CHOICE

• “GS” answer type extraction
  – UTS maps GS answers to “GS” types
  – No types found for 82 out of 406 questions

• Learning
  – Lemma, begins with “do” or “be”, contains “or”, contains digits, semantic type (using MetaMap), dependency label
  – Multi-class classification via Logistic regression (10-fold cross prediction)
Candidate Answer Generation

• Concepts
  – MetaMap annotated, LingPipe NER identified, OpenNLP chunker annotated NP and NP-PP-NP

• QUANTITY
  – POS tag of CD as the key token
  – Expansion: 3.0 -> 3.0 mm

• CHOICE
  – Head token of the “or” token as the first option
  – All children of the first option that have a dep-label of conj as alternative options.
  – Expansion

• CRF-based answer phrase
Candidate Answer Scoring

• Extend the approach used by Weissenborn et al. to 11 groups of features
  – Type coercion
  – CAO (candidate answer occurrence) count
  – Name count
  – Avg. covered token count
  – Stopword count
  – Token/concept overlap count
  – Token/concept proximity
  – LAT count
  – Parse proximity

• Use Logistic regression to learn the scoring function
Answer Pruning (for List Questions)

- Batch 4: an absolute threshold
- Batch 5: a relative threshold
- Collective reranking of candidates
Results (Phase B) - Tentative

- Exact answers

<table>
<thead>
<tr>
<th>Batch</th>
<th>Factoid</th>
<th>List</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Strict Accuracy</td>
<td>Lenient Accuracy</td>
</tr>
<tr>
<td>3</td>
<td>.1154 (1)</td>
<td>.2308 (1)</td>
</tr>
<tr>
<td>4</td>
<td>.4483 (1)</td>
<td>.6207 (1)</td>
</tr>
<tr>
<td>5</td>
<td>.2273 (1)</td>
<td>.3182 (1)</td>
</tr>
</tbody>
</table>
Error Analysis (Phase B)

• Concept type identification/answer type prediction (25)
• Concept identification (10)
  – Neurostimulation of which nucleus is used for treatment of dystonia?
    • Bilateral globus pallidus internus (Gpi)
• Complex answer (9)
  – “Effect”, “role”, “function”, etc.
  – What is the function of caspases?
    • Executors/mediators of apoptosis
• Mistakenly use question phrase as answer (7)
  – What is the effect of enamel matrix derivative on pulp regeneration?
    • EMD
Error Analysis (Phase B, cont’d)

• Tokenization (6)
  – t(11;22)(q24;q12)

• Definition question (3)
  – What is Piebaldism?
  – How are ultraconserved elements called when they form clusters?

• Question type (2)
  – Alpha-spectrin and beta-spectrin subunits form parallel or antiparallel heterodimers?
  – What is the risk of developing acute myelogenous leukemia in Fanconi?

• Snippets that have no information (2)
  – What is the main role of Ctf4 in dna replication?
    • Ctf4 remains a central player in DNA replication
Conclusion

• We present a three-layered architecture and the describe the components.
• The official evaluation results show the effectiveness of the proposed approach in factoid and list QA.
• We have been adopting BioASQ task and the benchmark in a number of CMU courses since 2014
  – 11-791 Design & Engineering of Intelligent Information Systems (final project): 30–70 students / year
  – 11-796/7 Question Answering (an option*)
  – 11-632 Data Science Analytics Capstone (an option*)
    * Other options include Entrance Exam World History task.
Conclusion (cont’d)

• Collaboration with Roche Pharmaceuticals
  – Funding for open source software development in biomedical question answering via Apache license
  – Internal use / development based on the open source release, but adapted to specific scenarios / proprietary resources

• In the stage of code refactoring for the purposes of
  – Open source release as the agreement requires
  – 2015 fall students in 11-791 for their final project
Thank you!

Zi Yang
PhD Candidate
Language Technologies Institute
School of Computer Science
Carnegie Mellon University

ziy@cs.cmu.edu
Retrieval Approaches (Phase A)

• **Document retrieval**
  – Lucene index (10K documents)
  – Negative Query Generation model (100 documents)
  – LETOR with random forest (10 documents)

• **Snippet retrieval**
  – Sentences as candidate snippets
  – Lucene, logistic regression for reranking

• **Concept retrieval**
  – MetaMap, LingPipe (GeneTag)
  – GoPubMed service

• **Triple retrieval**
  – Append [obj] and [sub] to each keyword
  – Enumerate all letter case possibilities
  – GoPubMed service
Results (Phase A)

- **Document retrieval**

<table>
<thead>
<tr>
<th>Batch</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>MAP</th>
<th>GMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.2310 (15)</td>
<td>.3242 (15)</td>
<td>.2311 (15)</td>
<td>.1654 (15)</td>
<td>.0136 (15)</td>
</tr>
<tr>
<td>4</td>
<td>.2144 (15)</td>
<td>.3320 (15)</td>
<td>.2263 (15)</td>
<td>.1524 (15)</td>
<td>.0081 (14)</td>
</tr>
<tr>
<td>5</td>
<td>.2130 (15)</td>
<td>.4474 (15)</td>
<td>.2605 (15)</td>
<td>.1569 (15)</td>
<td>.0267 (8)</td>
</tr>
</tbody>
</table>

- **Snippet retrieval**

<table>
<thead>
<tr>
<th>Batch</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>MAP</th>
<th>GMAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.1133 (3)</td>
<td>.1044 (5)</td>
<td>.0891 (3)</td>
<td>.0892 (1)</td>
<td>.0013 (5)</td>
</tr>
<tr>
<td>4</td>
<td>.1418 (5)</td>
<td>.1264 (10)</td>
<td>.1153 (8)</td>
<td>.0957 (5)</td>
<td>.0027 (6)</td>
</tr>
<tr>
<td>5</td>
<td>.1472 (9)</td>
<td>.1756 (9)</td>
<td>.1391 (9)</td>
<td>.1027 (9)</td>
<td>.0040 (5)</td>
</tr>
</tbody>
</table>