BIOMEDICAL QUESTION ANSWERING

by

Abhishek Bhandwaldar

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Computer Science

Charlotte

2017

Approved by:

____

Dr. Wlodek Zadrozny

____

Dr. Zbigniew W. Ras

____

Dr. Samira Shaikh

ABSTRACT

ABHISHEK BHANDWALDAR. Biomedical Question Answering. (Under the direction of DR. WLODEK ZADROZNY)

A Biomedical Question Answering system allows biomedical experts to use unstructured knowledge in a more effective way. Our system can index vast majority of biomedical articles and can do information retrieval to find most relevant articles to answer given an input question. The system can also summarize and return answer given a set of snippets from the extracted articles. The system can also answer in exact words and list of words depending on the type of question. We have evaluated our system on BioASQ dataset and have shown our results.

ACKNOWLEDGEMENTS

I would like to thank my guide and chair Prof. Wlodek Zadrozny for guiding me through this research study. I would also like to thank Prof. Zbigniew W. Ras for agreeing to be on my thesis committee. I also thank Prof. Samira Shaikh for being my graduate advisor and agreeing to be on my thesis committee. I would also like to acknowledge the assistance of Walid Shalaby for helping with query processing techniques and Hossein Hematialam for helping with biomedical entity mapping.

TABLE OF CONTENTS

LIST OF FIGURES vii

LIST OF TABLES viii

LIST OF ABBREVIATIONS 1

CHAPTER 1: INTRODUCTION 1

1.1. Problem Statement 1

1.2. Bioasq Competition 2

1.3. Related Work 2

CHAPTER 2: BioASQ Competition 5

2.1. Competition Format 5

2.1.1. Phase A Information Retrieval 5

2.1.2. Phase B Question Answering 8

2.2. Dataset 9

2.2.1. Training Dataset 9

2.2.2. Phase A Dataset 10

2.2.3. Phase B Dataset 11

CHAPTER 3: Evaluation 12

3.1. Phase B Evaluation 12

3.2. Phase 2 Evaluation 13

3.2.1. Yes/NO Question 13

3.2.2. Factoid question 13

3.2.3. List type question 14

3.2.4. Evaluating Ideal answers and Summary based questions 14

3.3. BioASQ Oracle 15

CHAPTER 4: Information Retrieval 16

4.1. Article Indexing 16

4.1.1. Query Processing 16

4.2. Results 16

CHAPTER 5: Question Answering 19

5.1. Summary based question answering 19

5.1.1. Results 20

5.2. Factoid Question Answering 21

5.2.1. Results 22

CHAPTER 6: CONCLUSIONS 24

REFERENCES 25

LIST OF FIGURES

FIGURE 2.1: Designated data sources in Disease ontology 6

FIGURE 2.2: Designated data sources in Gene ontology 7

FIGURE 2.3: Designated data sources in Uniprot 7

FIGURE 2.4: Designated data sources in Linked Life Data 7

FIGURE 2.5: Designated data sources in MESH 8

FIGURE 2.6: Designated data sources in UMLS 8

FIGURE 2.7: Example of required answers 9

FIGURE 2.8: Training dataset format 10

FIGURE 2.9: Phase A Dataset format 11

FIGURE 2.10: Phase B dataset format 11

FIGURE 5.1: Summarization pipeline 20

FIGURE 5.2: Factoid and list type question pipeline 22

LIST OF TABLES

TABLE 1.1: Summary of different approaches of participating systems. 4

TABLE 4.1: Result for document retrieval of phase A of BioASQ task 5b. 17

TABLE 4.2: Result for concept retrieval for phase A of BioASQ task 5b. 17

TABLE 4.3: Document retreival result on unchanged golden standard set. 17

TABLE 4.4: Document retrieval results of participating systems 18

TABLE 5.1: Results of Ideal answers 21

TABLE 5.2: Result of Factoid and List type Questions 23

CHAPTER 1: INTRODUCTION

A question answering system allows us to access and use information present in the unstructured form such as articles and documents. A well known example of QA system is the IBM Watson, which won jeopardy! challenge in 2011. This kind of systems are resourceful and find application in various natural language processing systems like chatbots, search engines etc. A Biomedical Question Answering is a system capable of answering a biomedical question using documents of that domain. Hence, aiding the biomedical experts in their research. The aim of this study is to build such system. We use the BioASQ dataset and participate in their competition to compare our system against other participating systems from various universities.

1.1 Problem Statement

BioASQ Task on Biomedical Semantic QA consists of using benchmark dataset containing train and test question along with the gold standard answers to develop a system that can respond with relevant concepts (from designated terminology and ontology), snippets (from retrieved articles), articles(from designated article repository), and answer the question. The task is comprised of two Phases. In phase A, questions are released and participants are required to do information retrieval to return most relevant articles, concepts, and snippets. In phase B, the released questions are then enhanced with golden standard articles, concepts, and snippets and participants are to respond with an exact answer, as well as summaries in natural language(dubbed as Ideal answer).

1.2 Bioasq Competition

BioASQ organizes challenges on biomedical semantic indexing and question answering (QA). The challenges include tasks relevant to hierarchical text classification, machine learning, information retrieval, QA from structured and unstructured data, multi-document summarization and many other areas. BioASQ is funded by the European Commission's Seventh Framework Programme and supported by National Library of Medicine of the National Institutes of Health. The competition is organized by Georgios Paliouras (NCSR "Demokritos", Greece and University of Houston, USA), Prof. Ioannis A. Kakadiaris (University of Houston, USA), and Anastasia Krithara (NCSR "Demokritos", Greece). BioASQ official website[1] describes its vision as to build an information system that can use the Biomedical knowledge dispersed in hundreds of heterogeneous knowledge sources and databases and aid Biomedical experts in their research.The challenge as per the website runs in two stages, designed to

1. adapt traditional semantic indexing and QA methods to the needs of biomedical experts, and
2. collect feedback and improve the experimental setting itself.

BioASQ aims to make progress in area of semantic indexing and question answering to make it more acceptable to the Biomedical community.

1.3 Related Work

The BioASA task 5b saw many approaches. The "Basic QA pipeline" describes phase A retrieval system by using Metamap for entity extraction, query expansion, and BM25 model for semantic similarity. For phase B they used the same technique except for the added step of removing the stop words. Schulze et al.[1] describes the system "HPI" for both phases based on the use of UMLS[2] and NER for entity

[1] http://www.bioasq.org/about/vision

identification and query formulation for document retrieval. For answering ideal answer they use LexRank summarization. The "USTB"[3] system which participated in phase A uses multiple query processing like the sequence dependence model and pseudo-relevance feedback ranking for document and snippet retrieval. The "fdu" system uses various query processing techniques like the query term weighting and pseudo-relevance feedback. "MQU" describes system using deep learning with regression and word embedding to generate the ideal answer for phase B. The "OAQA"[4] system used extractive summarization technique for answering ideal answers for phase B. They have used various algorithms including agglomerative clustering, Maximum Marginal Relevance, and sentence compression Wiese et al.[5] describe end to end neural network approach which is based on FastQA ([6]). The neural network takes a question and snippets and outputs start and end pointer in the snippet. Mourad et al.[7] describe their system which uses sentiment analysis for Yes/No question, entity identification using MetaMap and ranking them based on their frequency for List and Factoid questions. For generating Ideal answers they preprocess the snippets, rank them using BM25 model and concatenate top two. Below Table1.1 summarizes various techniques used by participating systems.

Table 1.1: Summarizes approaches taken by various participant system and the phase in which they participated. The approaches taken by system range from statistical based methods like the LexRank to machine learning based models to deep learning based models like the DeepQA, and FastQA. Table source[8]

Systems	Phase	Approach
Basic QA pipeline	A, B	MetaMap, BM25
Olelo	A, B	NER, UMLS, SAP HANA, SRL
USTB	A	sequential dependence models, ensembles
fdu	A	MESHLabeler,Language model, word similarity
UNCC	A	Stanford Parser, Semantic Indexing
MQU	B	deep learning, neural nets, regression
Oaqa	B	agglomerative clustering, tf-idf, word embeddings, maximum margin relevance
LabZhu	B	PubTator, Standford POS tool, ranking DeepQA B FastQA, SQuAD
sarrouti	B	UMLS, BM25, dictionaries

CHAPTER 2: BioASQ Competition

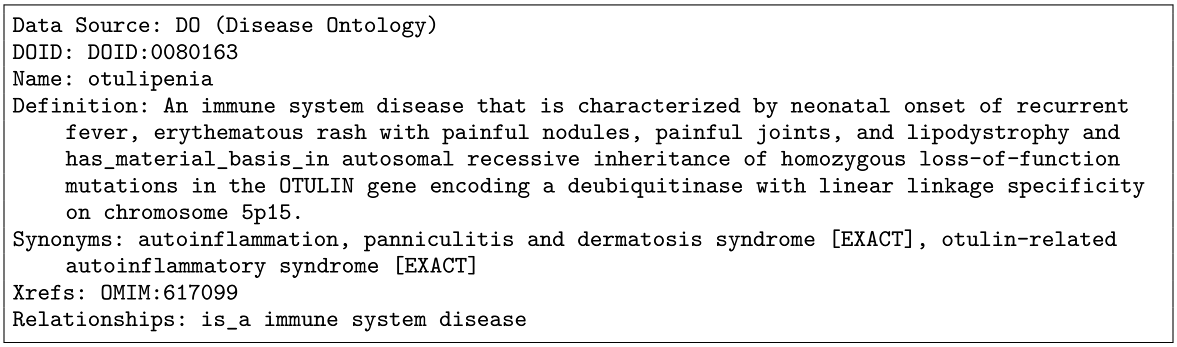
We focus on BioASQ semantic QA task. In this chapter we first go through the competition format, its dataset format, and its evaluation metric used for ranking participant submission.

2.1 Competition Format

The competition is conducted in two Phases. Phase A deals with the information retrieval part while Phase B deals with question answering. The competition makes available the training dataset for building and testing the system. Test data is released in 5 independent batches and two phases with the time gap of 24 hours. Each test batch set is released in the interval of one week and consist of 100 questions. The test batch for Phase A consist of questions and in Phase B dataset the same question are enhanced with the golden standard article set, concepts, and snippets from these articles. The submission is then ranked based on an ordered metric. Initial ranking is only indicative and the final ranking changes after biomedical experts go through each team's submission manually and augment the golden standard dataset with any entries that might seem very relevant but were missing. Previous year test batch sets are also available and the evaluation can be done through Oracle (BioASQ evaluation service for the older dataset.) Figure 2.9 shows the dataset format for phase A while figure 2.10 shows the dataset format of phase B.

2.1.1 Phase A Information Retrieval

In Information retrieval phase, participants are required to return the list of articles, concepts, snippets from the extracted article and RDF triples. The designated repository for the articles is the annual Baseline Medline February repository. The articles abstract and title is only considered. The article list needs to be for every question, ordered in decreasing confidence and at most 10 lengths. For concept extraction, the designated sources are the Disease Ontology (DO)[9], the Gene Ontology (GO)[10], the Joint Chemical dictionary[11], the Medical Subject Headings (MeSH), and the Universal Protein Resource (UniProt, SwissProt subset)[12]. For each question single list of concepts ordered in decreasing confidence and at the most length of 10 needs to be returned. The RDF triples are to be returned from the Linked Life data project. The Linked Life Data platform is a data warehouse that syndicates large volumes of heterogeneous biomedical knowledge in a common data model[1].Figure 2.4 shows a typical entry in Linked Life data repository. At most 10 RDF triples can be returned for each question ordered by decreasing by confidence. The snippets are to be retrieved from the retrieved documents. Each snippet will contain the unique identifier of the article from which it came, the character start and end location and the section from which it came from. A single snippets list needs to be returned for each question ordered by decreasing confidence and at most 10 can be returned.



```
Data Source: DO (Disease Ontology)
DOID: DOID:0080163
Name: otulipenia
Definition: An immune system disease that is characterized by neonatal onset of recurrent
    fever, erythematous rash with painful nodules, painful joints, and lipodystrophy and
    has_material_basis_in autosomal recessive inheritance of homozygous loss-of-function
    mutations in the OTULIN gene encoding a deubiquitinase with linear linkage specificity
    on chromosome 5p15.
Synonyms: autoinflammation, panniculitis and dermatosis syndrome [EXACT], otulin-related
    autoinflammatory syndrome [EXACT]
Xrefs: OMIM:617099
Relationships: is_a immune system disease
```

Figure 2.1: Example of data entry in Disease Ontology data source. Each entry has unique identifier or DOID to identify disease and is also used in generating response.

[1] http://participants-area.bioasq.org/general_information/Task5b/

```
Data Source: GO (Gene Ontology)
Accession: GO:0016757
Name: transferase activity, transferring glycosyl groups
Ontology: molecular_function
Synonyms: glycosyltransferase activity, transglycosidase activity, transglycosylase
    activity, transferase activity, transferring other glycosyl groups
Alternate IDs: GO:0016932
Definition: Catalysis of the transfer of a glycosyl group from one compound (donor) to
    another (acceptor). Source: GOC:jl, ISBN:0198506732
Comment
Subset: goslim_chembl, goslim_generic, gosubset_prok, goslim_yeast
```

Figure 2.2: Example of Gene Ontology data source. Each entry has ID which is used for identifying gene and also to generate response.

```
Data Source: Uniprot
ID: GL8D1_DANRE
Protein names : Recommended name:Glycosyltransferase 8 domain-containing protein 1 (EC
    :2.4.1.-)
Gene names Name:glt8d1 ORF Names:zgc:103525
Organism : Danio rerio (Zebrafish) (Brachydanio rerio)
Taxonomic identifier : 7955 [NCBI]
Taxonomic lineage : Eukaryota ? Metazoa ? Chordata ? Craniata ? Vertebrata ? Euteleostomi ?
     Actinopterygii ? Neopterygii ? Teleostei ? Ostariophysi ? Cypriniformes ? Cyprinidae ?
     Danio
Proteomesi
UP000000437 Componenti: Unplaced
```

Figure 2.3: Example of data entry in Uniprot. The uniques identifier or ID is used for generating response

```
Data Source: Linked Life Data
Name: CCNF, cyclin F
Organism: Homo sapiens ( Human ) - taxonomy: 9606
Gene Id: 899
Gene Type: protein-coding
Alternative label(s): FBX1, FBXO1, HGNC:1591, MIM:600227, Ensembl:ENSG00000162063, HPRD
    :02574, Vega:OTTHUMG00000128858
Chromosome: 16
Locations: 16p13.3
mRNAs: RefSeq: NM_001761.2
Proteins: RefSeq: NP_001752.2
Markers: UniSTS:156601, UniSTS:86758, UniSTS:84730
```

Figure 2.4: Example of data entry in Linked Life data. This data source is used to generate the RDF triples.

```
Data Source: MESH (Medical Subject Headings)
MeSH Heading: Scavenger Receptors, Class F
Tree Number(s): D12.776.543.750.705.940.742, D12.776.543.750.710.450.750.742
Unique ID: D051128
Scope Note: A group of structurally related scavenger receptors expressed predominately by
    ENDOTHELIAL CELLS. They-contain repeats of EPIDERMAL GROWTH FACTOR-like cysteine-rich
    motifs in their extracellular domains.
Entry Term(s): SR-F Proteins Registry Number: 0
```

Figure 2.5: Example of data entry in MESH. Each term or concept has MESH ID or Unique ID which is returned in response JSON object.

```
Data Source: UMLS (Unified Medical Language System)
Concept: [C0031453] Phenylalanine
Semantic Types: Amino Acid, Peptide, or Protein [T116], Biologically Active Substance [T123
    ], Pharmacologic Substance [T121]
Definitions
CSP | essential aromatic amino acid that is a precursor of melanin, dopamine,
    norepinephrine and thyroxine.
MSH | An essential aromatic amino acid that is a precursor of MELANIN; DOPAMINE;
    noradrenalin (NOREPINEPHRINE), and THYROXINE.
NCI | An essential aromatic amino acid in humans (provided by food), Phenylalanine plays a
    key role in the biosynthesis of other amino acids and is important in the structure and
     function of many proteins and enzymes. Phenylalanine is converted to tyrosine, used in
     the biosynthesis of dopamine and norepinephrine neurotransmitters. The L-form of
    Phenylalanine is incorporated into proteins, while the D-form acts as a painkiller.
    Absorption of ultraviolet radiation by Phenylalanine is used to quantify protein
    amounts. (NCI04)
```

Figure 2.6: Examples of data in UMLS[2]. We use this data source during question answering to get the semantic type and preffred name or biomedical concepts. We use Metamap which maps input string to UMLS meta thesaurus.

2.1.2 Phase B Question Answering

In the Question answering phase, the participants are expected to return the exact answer and the ideal answer. For the Yes/No question type the exact answer is either 'Yes' or 'No'. For the factoid type questions, the exact answer is a list of at most 5 entity names ordered by decreasing confidence. For the list type questions, each participating system is required to return a list of entity names of size no more than 100 and at most 100 characters wide each. No exact answer will be returned for summary type question. For each question (yes/no, factoid, list, summary), each participating system of Phase B may also return an ideal answer, i.e., a single paragraph-sized text ideally summarizing the most relevant information from concepts, articles, snippets, and triples retrieved in Phase A. Each returned "ideal" answer is intended to approximate a short text that a biomedical expert would write to answer the corresponding question (e.g., including prominent supportive information), whereas the "exact" answers are only "yes"/"no" responses, entity names or similar short expressions, or lists of entity names and similar short expressions; and there are no "exact" answers in the case of summary questions. The maximum allowed length of each "ideal" answer is 200 words[2]. Figure 2.7 shows example of answers required for phase B.

```
Summary Based Question Answer
       "ideal_answer": "The Yamanaka factors are the OCT4, SOX2, MYC, and KLF4
           transcription factors"

Yes/No Question
       "exact_answer": "yes"

Factoid(List type answers are similar to factoid except maximum allowed list lenght is 100
    rather than 5)
"exact_answer": [
                     ["Glutamine"],
                     ["ATXN3 gene"],
                     ["Machado-Joseph Disease"],
                     ["Proteins"],
                     ["Enzymes"]
              ],
```

Figure 2.7: Example answer for summary, factoid, list and yesno based questions

2.2 Dataset

The dataset is in JSON format. There are 5 test batches and each one contains 100 question. The dataset is provided in two parts, each one for each Phase. BioASQ also provide training dataset which contains 1799 questions. The test set from older competitions are also available for training and can be used.

2.2.1 Training Dataset

The training dataset[13] which is in JSON format contains the question and the gold standard articles, concepts, snippets, and RDF triples.

[2] http://participants-area.bioasq.org/general_information/Task5b/

```
{"questions":[
        {
                "type":"factoid",
                "body":"Is Rheumatoid Arthritis more common in men or women?",
                "id":"5118dd1305c10fae750000010",
                "ideal_answer": "Disease patterns in RA vary between the sexes; the condition
                     is more commonly seen in women,
                                                who exhibit a more aggressive disease and a
                                                    poorer long-term outcome.",
                "exact_answer": [
                                                        ["Women"]
                                                ],
                "documents": [
                                                "http://www.ncbi.nlm.nih.gov/pubmed/12723987"
                                                , ...
                                        ],
                "snippets":[
                                        {
                                                "document": "http://www.ncbi.nlm.nih.gov/pubmed
                                                    /22853635",
                                                "text": "The expression and clinical course of
                                                    RA are ...",
                                                "offsetInBeginSection": 559,
                                                "offsetInEndSection": 718,
                                                "beginSection": "sections.0"
                                                "endSection": "sections.0"
                                        }, ...
                                ],
                "concepts":[
                                                "http://www.diseaseontology.org/api/metadata/
                                                    DOID:7148", ...
                                        ],
                "triples": [
                                        {
                                                "s": "http://linkedlifedata.com/resource/umls/id
                                                    /C2827401",
                                                "p": "http://www.w3.org/2008/05/skos-xl#
                                                    prefLabel",
                                                "o": "http://linkedlifedata.com/resource/umls/
                                                    label/A17680439"
                                        },...
                                        ]
        }, ...
]}
```

Figure 2.8: The training data is in JSON format and contains the list of questions. Each question has question type, question id, question body, golden standard list of documents, snippets, concepts, and triples.

2.2.2 Phase A Dataset

The Phase A dataset contains the question, type and ID. The ID is used while returning the solution. Below is the dataset format.

```
{
       "questions": [
                            {
                                   "body": "Which two drugs are included in the
                                       Harvoni pill?",
                                   "type": "list",
                                   "id": "5896deff78275d0c4a000013"
                            }, ...
                     ]
}
```

Figure 2.9: Phase A dataset is in JSON format and contains the list of questions. Each question has question body or question text, question type namely list, factoid, yesno, or summary and question id used while returning response.

2.2.3 Phase B Dataset

The Phase 2 dataset contains the question, type, gold standard list of articles, and snippets.

```
{
"questions": [
              {
                     "body": "Which two drugs are included in the Harvoni pill?",
                     "documents": [
                                   "http://www.ncbi.nlm.nih.gov/pubmed/27276081", ...
                     ],
                     "type": "list",
                     "id": "5896deff78275d0c4a000013",
                     "snippets": [
                                   {
                                          "offsetInBeginSection": 0,
                                          "offsetInEndSection": 123,
                                          "text": "Will Sofosbuvir/Ledipasvir (Harvoni) Be
                                               Cost-Effective and Affordable for Chinese
                                              Patients Infected with Hepatitis C Virus?",
                                          "beginSection": "title",
                                          "document": "http://www.ncbi.nlm.nih.gov/pubmed
                                              /27276081",
                                          "endSection": "title"
                                          }, ...
                     ]
              }, ...
       ]
}
```

Figure 2.10: Phase B dataset is in JSON format and contains the list of questions. Each question contains question body, question type which specifies the type of response expected namely list, factoid, yesno, or summary, id which is required while generating response, and a list of golden documents, and snippets

CHAPTER 3: Evaluation

3.1 Phase B Evaluation

This phase is evaluated on unordered and ordered retrieval measures. The unordered retrieval measures used are the Mean, Precision and the F-measure for the article, concepts, snippets and RDF triples. The Mean and Precision for articles, concepts, and RDF triples, where TP represents True positive, FP represents false positive and FN represents false negative, is defined as

$$P = \frac{TP}{TP + FP} \tag{3.1}$$

$$R = \frac{TP}{TP + FN} \tag{3.2}$$

For snippets where G is ground truth snippet and S is the retrieved snippet Mean and Precision is defined as

$$P = \frac{|S \cap G|}{|S|} \tag{3.3}$$

$$R = \frac{|S \cap G|}{|G|} \tag{3.4}$$

The F-measure used is defined as

$$F = 2.\frac{P.R}{P + R} \tag{3.5}$$

The ordered retrieval measures used is the Mean average precision(MAP) and Geometric mean average precision (GMAP). The average precision is defined as

$$AP = \frac{\sum_{|L|}^{r=1} P(r).rel(r)}{|L_r|} \tag{3.6}$$

Where $|L|$ is number of items in list, $|L_r|$ is the number of relevant item, $P(r)$ is the precision of the first r elements in list, and $rel(r)$ equals to 1 if the r-th item in list is in the golden standard set. The MAP score is calculated by averaging all AP over set of questions $q_1, ..., q_n$, and is given by

$$MAP = \frac{1}{n} . \sum_{i=1}^{n} AP_i \tag{3.7}$$

where AP_i is the average precision of question q_i. The geometric mean precision is defined as geometric mean over average precision AP and is given as

$$GMAP = \sqrt[n]{\prod_{i=1}^{n} (AP_i + \epsilon)} \tag{3.8}$$

where ϵ is small number added to handle cased where $AP_i = 0$.

3.2 Phase 2 Evaluation

The exact answer in Phase B is evaluated as follow:

3.2.1 Yes/NO Question

For each Yes/No question the answer is compared to the golden standard answer and the accuracy Acc is computed as

$$Acc = \frac{c}{n} \tag{3.9}$$

where n is the number of yes/no questions and s is the number of correctly answered yes/no questions.

3.2.2 Factoid question

For factoid type question a list of at most 5 entity names is to be returned, arranged in decreasing confidence. Two types of score are calculated the strict accuracy (SAcc) and lenient accuracy (LAcc) and is defined as

$$SAcc = \frac{c_1}{n} \tag{3.10}$$

$$LAcc = \frac{c_5}{n} \tag{3.11}$$

where n is the number of factoid type question, c_1 is the number of factoid question answered correctly when first element of returned list is considered, and c_5 is the number of question answered correctly in lenient sense. Mean reciprocal rank (MRR) is also measured and is defined as

$$MRR = \frac{1}{n} . \sum_{i=1}^{n} \frac{1}{r(i)} \tag{3.12}$$

where $r(i)$ is the position of the topmost entity in returned list that matches the entity name in golden standard set.

3.2.3 List type question

For list type question the answer is a list of at most 100 entity names or short phrases.This list is compared to golden standard and Precision, recall, and F-measure is calculated for each question. By averaging the precision, recall and F-measure we get mean average precision, mean average recall, and mean average F-measure score.

3.2.4 Evaluating Ideal answers and Summary based questions

For each (yes/no, factoid, summary), we are required to return single paragraph sized summarization of the retrieved relevant information from the article, concepts and snippets. The answers are evaluated manually by biomdecal experts and automatically. For automatic evaluation two scores are calculate namely the *ROUGE-N* and *ROUGE-S* score. The *ROUGE-N* score which computes the overlap between constructed summary/ideal answer by a system S and a set $Refs$ of reference summaries using n-grams is defined as

$$ROUGE\text{-}N(S|Refs) = \frac{\sum_{R \in Refs} \sum_{g_n \in R} C(g_n, S, R)}{\sum_{R \in Refs} \sum_{g_n \in R} C(g_n, R)} \tag{3.13}$$

In above definition g_n is a word n-gram, $C(g_n, S, R)$ is the number of times g_n co-occurs in S and a reference summary R, and $C(g_n, R)$ is the number of times g_n occurs in reference R. The *ROUGE-S* score uses skip bigrams, instead of n-grams, when computing the overlap. *ROUGE-SU* is similar to *ROUGE-S* but it also counts the unigrams that occur in both S and $Refs$. Finally the *ROUGE-SU4* is a version of *ROUGE-SU4* with maximum distance between any skip bigram limited to 4. For evaluation BioASQ uses the *ROUGE-2* and *ROUGE-SU4* scores.

3.3 BioASQ Oracle

The BioASQ oracle[1] is where we can download old datasets as well as test our system's performance on these test sets. According to the official website the system can be tested by uploading the results for the appropriate test batch set and the evaluation usually takes few minutes.

[1] http://participants-area.bioasq.org/oracle/

CHAPTER 4: Information Retrieval

In phase A, for every question, we are required to retrieve and return articles, concepts, and snippets. In this chapter, we will go through the process of article retrieval.

4.1 Article Indexing

The designated article repository for this task is the Annual MedLine baseline repository for 2017 which contain approximately 26 million articles. We use Lucene for indexing, and the articles are indexed on abstract and title.

4.1.1 Query Processing

For every question, we retrieve a list of articles. To do this we first create the query by preprocessing the question. Preprocessing of the question involves passing the question through Stanford POS tagger[14], and then concatenating the Noun Phrases from the original question. We use this as our query preprocessing technique for BioASQ task 5b. We also query the designated sources using the noun phrases we identified in previous steps to gain more information like the synonyms, hyponyms/hypernyms relationships, and augment our original query.

4.2 Results

Table 4.1 shows results obtained by submitting our system in test batch 1,3, and 5 for document retrieval. From Table 4.4 it can be seen that our system performed well in the unordered measure namely the mean-precision, f1 measure, but suffered in the ordered measure score MAP and GMAP. Table 4.1 scores are calculated after the biomedical experts inspected the top k concepts, articles, snippets, and triples of each system, i.e., the k concepts, articles, snippets, and triples that each system is most confident about, in order to add to the corresponding golden sets any correct (relevant) items that the biomedical experts had missed, but the systems managed to retrieve. Table 4.3 shows the system score on unchanged golden dataset. The Table 4.2 shows result for Test batch 1, and 3 for concept retrieval.

Table 4.1: Result for document retrieval of phase A of BioASQ task 5b.

Test Batch	Mean precision	Recall	F-Measure	MAP	GMAP
Test Batch 1	0.3478	0.2698	0.2544	0.1840	0.0042
Test Batch 3	0.3301	0.2354	0.2172	0.1329	0.0022
Test Batch 5	0.3043	0.2496	0.2177	0.1157	0.0022

Table 4.2: Result for concept retrieval for phase A of BioASQ task 5b.

Test Batch	Mean precision	Recall	F-Measure	MAP	GMAP
Test Batch 1	0.3136	0.5862	0.3548	0.1543	0.0570
Test Batch 3	0.2022	0.3894	0.2456	0.1202	0.0104

Table 4.3: Result of document retrieval for Phase A of BioASQ task 5b on unchanged golden standard set. Note: The result for test batch 5 are missing as after the submission of test batch set 5 score on unchanged golden standard set where not available for much time

Test Batch	Mean precision	Recall	F-Measure	MAP	GMAP
Test Batch 1	0.2438	0.3523	0.2538	0.1080	0.0017
Test Batch 2	0.2317	0.3340	0.2322	0.0825	0.0009

Table 4.4: Document retrieval result of all participating systems from task 5b test batch set 3. As can be seen UNCC System 1 performed best on unordered measures, mean-precision and F1 measure but performed poorly on ordered measures like the GMAP and MAP. Table source [8]

System	Mean precision	Recall	F-Measure	MAP	GMAP
ustb-prir4	0.1707	0.4787	0.2200	0.1143	0.0066
ustb-prir1	0.1680	0.4750	0.2155	0.1108	0.0060
fdu2	0.1645	0.4628	0.2135	0.0976	0.0059
ustb-prir2	0.1737	0.4754	0.2220	0.1134	0.0059
ustb-prir3	0.1620	0.4803	0.2111	0.1157	0.0050
fdu	0.1615	0.4475	0.2120	0.1021	0.0049
testtext	0.1610	0.4690	0.2087	0.1138	0.0048
fdu4	0.1420	0.4310	0.1856	0.0926	0.0044
fdu3	0.1390	0.4098	0.1809	0.0976	0.0031
UNCC System 1	**0.2317**	**0.3340**	**0.2322**	**0.0825**	**0.0009**
fdu5	0.1060	0.2461	0.1298	0.0737	0.0007
Olelo	0.1327	0.2444	0.1481	0.0658	0.0005
HPI-S1	0.0823	0.2152	0.0997	0.0464	0.0005
KNU-SG	0.0730	0.2149	0.0967	0.0521	0.0005
c-e-50	0.0720	0.1921	0.0861	0.0547	0.0003
c-50	0.0720	0.1921	0.0861	0.0547	0.0003
c-idf-qe-1	0.0720	0.1921	0.0861	0.0547	0.0003
c-f-200	0.0720	0.1921	0.0861	0.0547	0.0003

CHAPTER 5: Question Answering

In this chapter, we will go through our approach for phase B of BioASQ task 5b question and answering. The various question types in this phase are the list type, factoid type, yes/no and summary type. We only discuss our approach for summary, factoid, and list type questions. We did not participate in yes/no question as dataset was highly skewed with some batches having almost all yes answers.

5.1 Summary based question answering

In summary based question we are given the golden standard documents and snippets. The answer expected is a paragraph-sized summary of the snippets of no more than 200 words. We use extractive summarization to generate the summary. The summarization pipeline is as follows:

1. Sentence segmentation of input snippets and removal of similar sentences arising due to overlapping snippets.

2. Extract biomedical entities and their semantic type from each sentence and input question. Tag each sentence and the question with the set of semantic types.

3. We compare semantic type set of question and every sentence and select sentence based on the maximum intersection.

4. Next, we augment the question set by doing the union of the question set with the set of the selected sentence and add the sentence to summarization.

5. We repeat above steps until either we run out of sentences or the summary size of 200 words is reached.

6. We concatenate all selected sentences to generate the final summary.

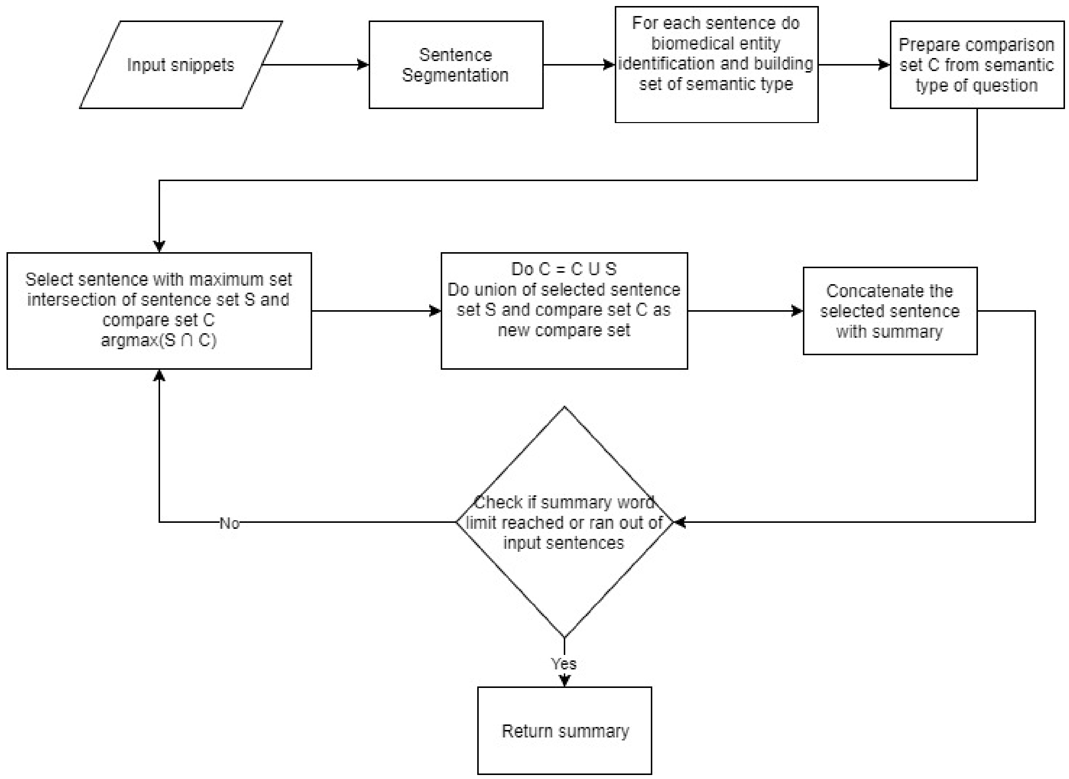

Input snippets
Sentence Segmentation
For each sentence do biomedical entity identification and building set of semantic type
Prepare comparison set C from semantic type of question
Select sentence with maximum set intersection of sentence set S and compare set C argmax(S ∩ C)
Do C = C U S
Do union of selected sentence set S and compare set C as new compare set
Concatenate the selected sentence with summary
Check if summary word limit reached or ran out of input sentences
No
Yes
Return summary

Figure 5.1: As shown in summarization pipeline the input is snippets and question. We use Metamap for biomedical entity identification. Output is summary such that its length does not exceed the 200 word limit.

5.1.1 Results

We used BioASQ oracle to test our system on previous years dataset and compare them with other submission for those batches. Our system got highest ROUGE-2 and ROUGE-SU4 score among all systems in every batch.

Table 5.1: Results of Ideal answers on all previous batch sets acquired using BioASQ oracle. The results are arranged from most recent to least. The table shows ROUGE-2 and ROUGE-SU4 scores, as discussed in 3.2.4, calculated for submitted answer.

Test Batch	Rouge-2	Rouge-SU4
Task 4B Batch 5	0.7347	0.7308
Task 4B Batch 4	0.7196	0.7177
Task 4B Batch 3	0.6364	0.6527
Task 4B Batch 2	0.6777	0.6897
Task 4B Batch 1	0.6918	0.7024
Task 3B Batch 5	0.5651	0.5672
Task 3B Batch 4	0.5848	0.5950
Task 3B Batch 3	0.5994	0.6128
Task 3B Batch 2	0.5451	0.5674
Task 3B Batch 1	0.5240	0.5368
Task 2B Batch 5	0.3967	0.4180
Task 2B Batch 4	0.4201	0.4458
Task 2B Batch 3	0.4731	0.4754
Task 2B Batch 2	0.4075	0.4258
Task 2B Batch 1	0.5313	0.5326
Task 1B Batch 2	0.3319	0.3596
Task 1B Batch 1	0.3032	0.3276

5.2 Factoid Question Answering

For factoid type and list type question our algorithm is as follow:

1. We first do sentence segmentation and identify all biomedical entities using Metamap.

2. Next, we calculate the score for each entity which is given as frequency + overlap

similarity of the sentence in which entity occurs with the question. The score takes into account the frequency of entity in all snippets and the similarity score of all the sentences in which the entity occurs and the question.

3. We also consider the IDF score of each word as this will prevent any common biomedical terms like cell, life, or science to be considered as answers. The IDF score used is calculated by indexing the Medline article repository using Lucene, which did in our phase A, and then getting the term frequency from the index. We multiply the IDF score with entity score to get final entity score.

4. Next, we rank the entities based on the score and return the top 5 as answers for factoid type and top 100 for list type question.

5. For list type question we return 100 entities answers.

Entire algorithm is summarized in figure 5.2.

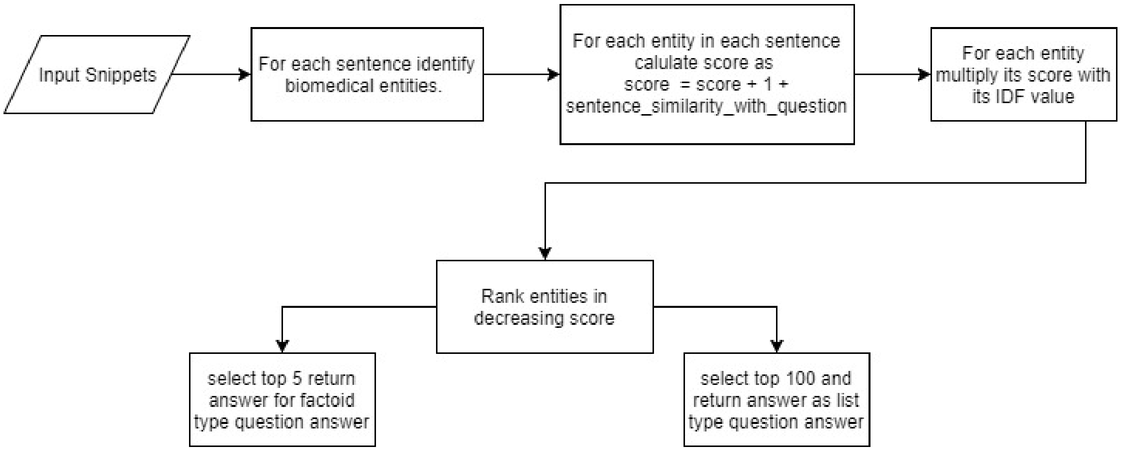

Input Snippets
For each sentence identify biomedical entities.
For each entity in each sentence calulate score as score = score + 1 + sentence_similarity_with_question
For each entity multiply its score with its IDF value
Rank entities in decreasing score
select top 5 return answer for factoid type question answer
select top 100 and return answer as list type question answer

Figure 5.2: Factoid and list type question pipeline. The input is list of snippets and the biomedical entity identification is done using Metamap and UMLS. Each entry in returned list is also a list of synonyms.

5.2.1 Results

Table 5.2 shows results for factoid and list type questions for our current technique.

Table 5.2: Result of factoid and list type questions evaluated on task 4b dataset.The table has SAcc and LAcc score as discussed in section 3.2.2, calculated for factoid type questions. Table also shows Mean precision, Recall, and F-measure calculated for List type questions, as discussed in section 3.2.3

Test Batch	Factoid SAcc	Factoid LAcc	Factoid MRR	List Mean Precision	List Recall	List F-measure
Task 4B Test Batch 5	0.0303	0.0909	0.0556	0.0316	0.2114	0.0536
Task 4B Test Batch 4	0.0323	0.0323	0.0323	0.0162	0.2717	0.0293
Task 4B Test Batch 3	0.0385	0.1538	0.0865	0.0191	0.2969	0.0341
Task 4B Test Batch 2	0.0323	0.0645	0.0403	0.0311	0.2525	0.0509
Task 4B Test Batch 1	0.0256	0.1026	0.0521	0.0051	0.0909	0.0096

CHAPTER 6: CONCLUSIONS

We created a biomedical question answering system which is able to find answer given question from unstructured data like the Medline article repository. Our system got the highest score for information retrieval for unordered measure compared to ordered measure. We conclude that our systems information retrieval results were relevant but needed to be re-ranked. Our system also got the highest score in summary based question answering. But our system did not perform well on factoid and list-based questions. This difference in performance is because our system is able to find the most relevant sentence to answer the question, as evidenced from summary answer performance, but is not able to efficiently extract the exact answer from that sentence. We also saw other approaches that are similar to the lexrank based techniques and other different approaches like the FastQA neural network model. The score of the wining system over the years has improved which shows the increasing quality of systems in the biomedical domain.

REFERENCES

[1] F. Schulze, R. SchÃŒler, T. Draeger, D. Dummer, A. Ernst, P. Flemming, C. Perscheid, and M. Neves, "Hpi question answering system in bioasq 2016," 01 2016.

[2] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, no. Database-Issue, pp. 267–270, 2004.

[3] Z. Jin, B. Zhang, F. Fang, L. Zhang, and X. Yin, "A multi-strategy query processing approach for biomedical question answering: Ustb_prir at bioasq 2017 task 5b," in Cohen *et al.* [7], pp. 373–380.

[4] Z. Yang, Y. Zhou, and E. Nyberg, "Learning to answer biomedical questions: Oaqa at bioasq 4b," 01 2016.

[5] G. Wiese, D. Weissenborn, and M. L. Neves, "Neural question answering at bioasq 5b," in Cohen *et al.* [7], pp. 76–79.

[6] D. Weissenborn, G. Wiese, and L. Seiffe, "Fastqa: A simple and efficient neural architecture for question answering," *CoRR*, vol. abs/1703.04816, 2017.

[7] K. B. Cohen, D. Demner-Fushman, S. Ananiadou, and J. Tsujii, eds., *BioNLP 2017, Vancouver, Canada, August 4, 2017*, Association for Computational Linguistics, 2017.

[8] A. Nentidis, K. Bougiatiotis, A. Krithara, G. Paliouras, and I. A. Kakadiaris, "Results of the fifth edition of the bioasq challenge," in Cohen *et al.* [7], pp. 48–57.

[9] L. M. Schriml, C. Arze, S. Nadendla, Y. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe, "Disease ontology: a backbone for disease semantic integration," *Nucleic Acids Research*, vol. 40, no. Database-Issue, pp. 940–946, 2012.

[10] M. Ashburner, "Gene ontology: Tool for the unification of biology," *Nature Genetics*, vol. 25, pp. 25–29, 2000.

[11] K. M. Hettne, R. H. Stierum, M. J. Schuemie, P. J. M. Hendriksen, B. J. A. Schijvenaars, E. M. van Mulligen, J. Kleinjans, and J. A. Kors, "A dictionary to identify small molecules and drugs in free text," *Bioinformatics*, vol. 25, no. 22, pp. 2983–2991, 2009.

[12] "The universal protein resource (uniprot) in 2010," *Nucleic Acids Research*, vol. 38, no. Database-Issue, pp. 142–148, 2010.

[13] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artières, A. Ngonga, N. Heino, É. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, and G. Paliouras, "An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition," *BMC Bioinformatics*, vol. 16, pp. 138:1–138:28, 2015.

[14] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003* (M. A. Hearst and M. Ostendorf, eds.), The Association for Computational Linguistics, 2003.