A Survey of Datasets for Biomedical Question Answering Systems

Muhammad Wasim, Dr. Waqar Mahmood, Dr. Usman Ghani Khan Al-
Khawarizmi Institute of Computer Science
University of Engineering and Technology, Lahore

***Abstract*—The massively ever increasing amount of textual and linked biomedical data available online poses many challenges for information seekers. So, the focus of information retrieval community has shifted to precise information retrieval, i.e. providing exact answer to a user question. In recent years, many datasets related to Biomedical Question Answering (BioQA) have emerged which the researchers can use to evaluate the performance of their systems. We reviewed these biomedical datasets and analyzed their characteristics. The survey in this paper covers these datasets for BioQA and has a two fold purpose: to provide an overview of the available datasets in this domain and to help researchers select the most suitable dataset for benchmarking their system.**

***Keywords*—Biomedical; QA system; review; survey**

I. Introduction

The massive amount of textual data on web makes finding information a challenging task. Although *Information Retrieval* (IR) systems are developed to cater this problem, such systems just provide a list of documents instead of precise information [1]. The information seeker thus needs to process the provided document list to filter the required information. To overcome this problem, the research and development in the area of question answering is in progress [2]–[4]. To accelerate research in this area, many research initiatives including Text Retrieval Conference(TREC)[1] have been taken to develop benchmark datasets for open domain question answering [5]–[12]. Early research focused on heuristic based methods such as patterns for specific question types or redundancy of data over web to answer questions mostly from textual documents [13], [14]. The current focus in question answering systems is to use statistical learning approaches and neural networks to answer questions [15]–[19] both in textual and linked data. Moreover, the trend in question answering is changing from open domain to restricted domains such as biomedical [20], [21]. Focusing on restricted domains makes finding solution to specific answer types easy as different patterns can be used to determine question and answer types and domain knowledge can be incorporated [22].

The biomedical data on web can be broadly divided into two main categories i.e. textual data and linked data [23]. Textual data includes scientific literature published by biomedical journals and may include other authentic websites on biomedical domain. Linked data provides information about how different biomedical entities are related to each other and provides mechanism to inference and derive new knowledge. Linked biomedical data is one of the most important types of data in linked data world as around one-tenth of all linked data available online is biomedical data [10]. It may include drugs, compounds, diseases, genomics, proteomics or other nomenclature/lexical ontologies. Fig. 1 depicts the general categorization of available biomedical data. Biomedical experts usually publish their research both in the form of research articles and submit their experimental results to linked data repositories. To validate their hypothesis during any further experimentation, biomedical experts need to consult both textual and linked data sources manually. Biomedical question answering systems should be able to derive such answer automatically by combining evidence from both textual and linked data sources.

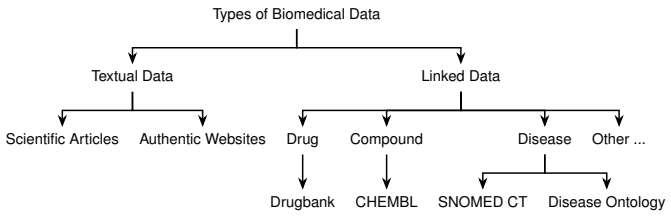

Fig. 1. Information sources in biomedical domain.

To tap the real potential of this heterogeneous data, techniques need to be developed and biomedical question answering datasets play an important role in this process as researchers can use it to evaluate the performance of their systems. Question answering datasets usually contain two main components: 1) *Information Need* or question in natural language and 2) the *corpus or linked data* which will help in answering the questions to fulfill some information need. Questions may also have additional information such as question types such as passage, factoid, list or summary. The focus of *passage* type question is to return relevant passages. The *factoid* type questions deal with single entity such as genes, proteins, disease, symptoms etc. List questions typically return list of factoids. The answer of *yes/no* type question is either yes or no. Finally, the aim of the *summary* question is to provide answer in a summarized form and may include definitional questions. Such questions typically start with *What is* phrase. There are also datasets available where question type is *multiple choice question*. The focus of such type of questions is to improve the passage comprehension and inferencing capabilities of question answering systems. Question types may help a QA system to focus on one particular answer

[1] http://trec.nist.gov/data/qa.html

strategy [24] and following are some typical questions asked in biomedical domain:

- Passage Question
- Factoid Question
- List Question
- Multiple Choice Question
- Yes/No Question
- Summary Question

Biomedical experts have certain information needs. They might be interested in knowing how genes interact with organ functions or what role some genes play in a particular disease.They might be further interested in *proteins*, *protein-protein interactions*, *mutations*, *drugs*, *adverse effects*, *cell or tissue types* and *signs or symptoms*. Other entities of interest might be *chemicals*, *species*, *pathways*, *genetic variations*, and *patient characteristics*. Heterogeneous biomedical sources are normally required to fulfill an information need as shown in Fig. 2. Identifying them early in the question answering pipeline helps narrow down candidate answers in later stages [24].

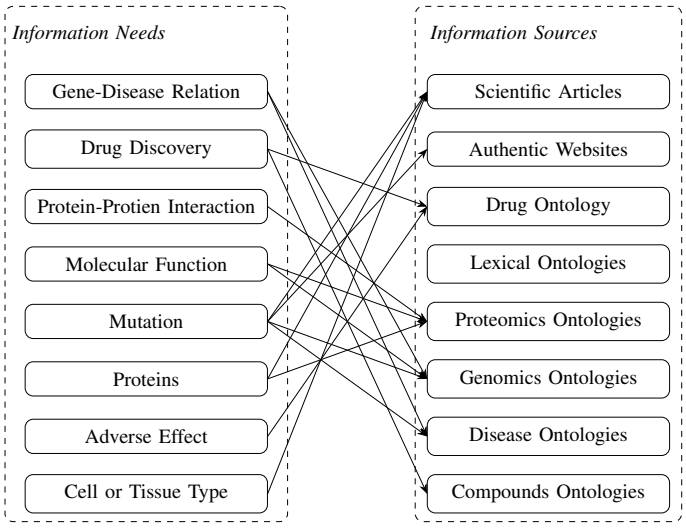

Fig. 2. Mapping biomedical information needs to sources.

This paper is organized as follows. Section II describes different biomedical question answering datasets and their specific characteristics. A comparative study of such datasets is presented in the next section and finally the paper is concluded in the last section.

II. Datasets for Biomedial Question Answering

This section enlists the biomedical question answering datasets in chronological order.

A. TREC Genomics Track

Genomics corpus provided by TREC was one of the pioneer datasets developed for Biomedical Question Answering. The dataset was prepared for TREC Genomics track held in 2006 and 2007 [20], [21]. The challenge was to retrieve relevant passages against a topic question from a corpus of 162,259 documents. The corpus was collected by crawling 49 biomedical journals and covered various biomedical categories. A set of 28 and 36 topic questions was assembled in 2006 and 2007 respectively. Some example questions from both collections are shown in Tables I and II. The biomedical entities covered in these two years included: *genes*, *proteins* and *gene mutations* in 2006; and *antibodies*, *cell or tissue type*, *disease*, *drug*, *gene*, *molecular function*, *mutations*, *pathways*, *proteins*, *symptoms*, *toxicities*, and *tumor types* in 2007.

TABLE I. TREC 2006 Genomics Track Topic Questions

Example Questions
How does L2 interact with L1 to form HPV11 viral capside?
How does p53 affect apoptosis?
How do mutations in the Pes gene affect cell growth?

TABLE II. TREC 2007 Genomics Track Topic Questions

Questions
What serum [PROTEINS] change expression in association with high disease activity in lupus?
What [MUTATIONS] in the Raf gene are associated with cancer?
What [DRUGS] are associated with lysosomal abnormalities in the nervous system?

The dataset was focused on scientific articles only and limited number of questions were not sufficient to evaluate the performance of large-scale QA system. Moreover, no annotations were provided with the corpus or questions so machine learning algorithms could not be effectively used.

B. Question Answering for Machine Reading Evaluation (QA4MRE): Biomedical Text about Alzheimer's Disease

QA4MRE[2] for biomedical data [25] differs from TREC datasets because the focus of the dataset is on passage comprehension and multiple answers are already provided with each question. Corpus is used as a *background knowledge* i.e. it is used to acquire knowledge instead of directly answering the question. Collection developed as a *background knowledge* for the task was named Alzheimer's Disease Literature Corpus (ADLC corpus) was collected from different sources: 66,222 pubmed abstracts were from *PubMed*, 8249 open access full articles from *pubmed central*, 1041 full text articles from *pubmed central* in HTML and text format. Moreover, 379 full text articles and 103 abstracts were added in this collection from a list of popular articles on Alzheimer disease. The test set is composed of four reading tests where each test consists of one document and 10 questions related to that document and five answer choices per question. Table III shows some sample questions from the collection. The systems had to either select answer from the given list of possible answers or leave the question unanswered. The background collection, test documents and questions were annotated with *word*, *lemma*, *chunk*, *part of speech*, *named entity*, *parent node in the dependency tree*, *dependency syntax label*, *UMLS entity name*, and *named entity*.

[2]http://celct.fbk.eu/QA4MRE/

TABLE III. QA4MR ABOUT BIOMEDICAL TEXT ON ALZHEIMER'S DISEASE

Questions	Options
Which CLU isoform has a consistently higher gene expression?	CLU2 ribosomal protein L13A CLU1 allele PNGase
Which hormone can control the expression of CLU isoforms?	real-time PCR cDNA AD rs11136000 androgen
What effect do androgens have on CLU2 gene expression?	association repression inhibition activation expression

These annotations were performed automatically using state of the art tools such as *GDep parser*, *CLiPS NE Tagger*, and *ABNER tagger*. The named entities considered for the task were *genes* and *gene products*, *chemicals*, *drugs*, *symptoms*, *experimental methods*, *species*, *pathway*, *cellular*, *genetic variation*, *adverse effect*, *dose*, *timing*, *patient characteristics* etc. The challenge was more about checking the inferencing capabilities of the system and multiple choice question proposed a new dimension of research. Moreover background collection was also different in the perspective. There was no retrieval sub-system and linked data was not used as a source in the dataset.

C. QALD-4: Biomedical Question Answering over Interlinked Data

Question Answering over linked data (QALD)[3] started in 2011 viewing the wide spread of linked data over the web and focused on converting information need into standard semantic query processing and inferencing. The objective of the dataset is to establish a standard against which question answering systems over structured data can be evaluated and compared. The fourth year competition had three tracks i.e. *multilingual question answering*, *biomedical question answering over interlinked data*, and *hybrid question answering*. Biomedical data was selected as there are many structured datasets available in this domain and answer to information need can be satisfied only if evidence is combined from multiple sources. Three biomedical datasets were combined for this competition i.e. 1) *SIDER*, which describes drugs and it's side effects; 2) *diseasome*, which provides information about diseases and genetic disorders; and 3) *drugbank*, which provides FDA-approved active compounds of medication. There were a total of 25 training questions and 25 similar test questions. Some example questions from the dataset are shown in Table IV.

There are very limited number of test and training questions and the only target of dataset is on three ontology sources. The questions were prepared in a manner that multiple sources needed to be inquired to find answer to a question. The systems using this dataset should address the problem of converting natural language questions to SPARQL queries efficiently.

[3]http://qald.sebastianwalter.org/index.php?x=home&q=1

TABLE IV. SOME QUESTIONS FROM QALD-4 BIOMEDICAL DATASET

Questions
Which genes are associated with Endothelin receptor type B?
Which genes are associated with subtypes of rickets?
Which drug has the highest number of side-effects?
List drugs that lead to strokes and arthrosis.
Which drugs have a water solubility of 2.78e-01 mg/mL?

D. BioASQ large-scale Biomedical Semantic Indexing and Question Answering

The focus of BioASQ challenge has been to assemble information from multiple heterogeneous sources in order to answer real-life biomedical experts' questions. The competition is being held every year since 2013. The challenge consists of two tasks 1) Biomedical semantic indexing; and 2) Semantic question answering: semantic answering is further divided into two tasks i.e. retrieval of relevant documents, snippets and triplets (Phase A) and finding the precise answer of the question (Phase B). The focus of the challenge is on drugs, targets, disease and covers both textual and linked data. The selected sources for these categories are shown in Table V.

TABLE V. THE RESOURCES THAT WERE INCLUDED IN BIOASQ CHALLENGE

Focus	Resources
Drugs	Joche
Targets	Gene Ontology, UniProt
Diseases	Disease Ontology
General Purpose	MeSH
Document Sources	PubMed, PubMed Central
Linked Data	LinkedLifeData

The challenge involves text/passage retrieval, RDF triplet retrieval, QA for exact answer, multi document summarization and natural language generation. The answer of any question may be factoid or passage depending on the type of question. The system could provide exact answer or ideal answer where ideal answer is paragraph-sized summary of the answer. The dataset contained development and test questions. There are four type of questions in dataset 1) Yes/No; 2) factoid; 3) list; and 4) summary. All questions expect exact and ideal answer except summary question where only ideal answer is expected. Table VI shows example questions for each type. The training and test questions for each year are: 29, 282 in 2013; 310, 500 in 2014; 810, 500 in 2015; 1307, 500 in 2016; and 1799, 500 in 2017 [26].

TABLE VI. QUESTIONS TYPES IN BIOASQ DATASET

Question Type	Example Question
Yes/No	Is miR-21 related to carcinogenesis?
Factoid	Which is the most common disease attributed to malfunction or absence of primary cilia?
List	Which human genes are more commonly related to craniosynostosis?
Summary	What is the mechanism of action of abiraterone?

III. DISCUSSION

Biomedical data is available in both textual and linked formats. Therefore, the question answering datasets should also provide good coverage of both of these sources. Fig. 3

shows the number of datasets which provide linked, textual and heterogeneous biomedical data. Only one dataset provides linked data source. Three datasets provide textual data out of which two use the same corpus with different question sets. The third textual dataset uses background collection as data source. The four datasets provided by BioASQ provide heterogeneous data. The dataset mostly targets open and publicly available textual and linked datasets and provides good number of training examples. Textual datasets can be searched using keywords based IR techniques. Linked data requires specific query format conversion supported by linked data repositories known as SPARQL. Heterogeneous datasets require both keyword based queries and linked data specific query formats. Moreover, to produce answer in natural language, answers from linked data are normally required to pass through natural language generation module. The datasets providing heterogeneous sources are most effective as both sources are equally important in the context of biomedical domain.

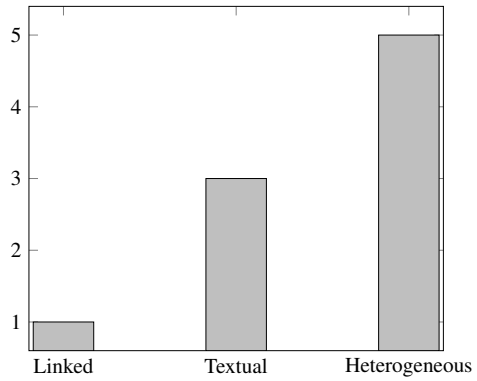

Fig. 3. Datasets with respect to type of data.

The number of questions provided with each dataset have increased over years. Initially, only 28 questions were provided in TREC Genomics track with no training set. Training question were first introduced in QALD-4 with 25 training questions. The trend from then has only increased and current year's BioASQ track contained 1307 training questions. This increase in information needs of real world biomedical experts' needs is essential to build practically usable QA systems. More training data may greatly aid in building machine learning based algorithms. Fig. 4 shows the number of training and test questions in all presented datasets.

Table VII shows a comparative table comparing all the datasets presented in this paper. Each dataset's strength and weakness are also mentioned to help researchers select appropriate dataset for their research. To summarize:

- If the QA system is to be built upon IR system, TREC dataset for 2006 and 2007 with a total of 64 questions provide a good starting point.
- To build and test a QA system which queries multiple linked data sources, QALD dataset can be used.
- State-of-the-art dataset is provided by the recent BioASQ challenge. The dataset can be used to evaluate:

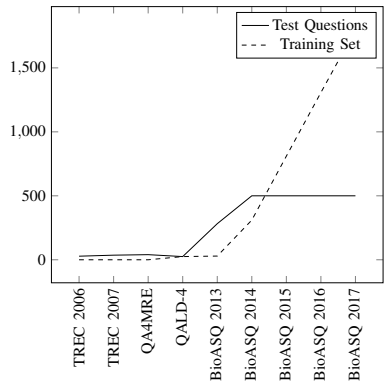

Fig. 4. Datasets - training vs test questions.

  - Performance of system on heterogeneous data sources.
  - Inferencing capabilities of a QA system.
  - Ability to generate natural language answer.
  - The quality of summarized answers.

IV. Conclusion

Question answering aims to provide a practical solution to the information overload problem. The availability of biomedical dataset highlights both the challenges and opportunities present in this domain. The availability of training data provides the opportunity to tailor systems in learning from examples. Moreover, the initiatives from BioASQ on heterogeneous dataset is paving the way towards better datasets to evaluate the effectiveness of biomedical QA systems on a larger scale. BioASQ provides the opportunity to exploit heterogeneous sources, perform inference, and produce summary for ideal answers. All the datasets have their strengths and weakness but overall BioASQ provides heterogeneous sources and more training data. In future, we shall investigate the systems (and their characteristics) which work best for each dataset.

References

[1] O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, L. Monceaux, I. Robba, and A. Vilnat, "Finding an answer based on the recognition of the question focus." in *TREC*, 2001.

[2] V. Lopez, V. Uren, M. Sabou, and E. Motta, "Is question answering fit for the semantic web?: a survey," *Semantic Web*, vol. 2, no. 2, pp. 125–155, 2011.

[3] S. K. Dwivedi and V. Singh, "Research and reviews in question answering system," *Procedia Technology*, vol. 10, pp. 417–424, 2013.

[4] G. Zayaraz *et al.*, "Concept relation extraction using naïve bayes classifier for ontology-based question answering systems," *Journal of King Saud University-Computer and Information Sciences*, vol. 27, no. 1, pp. 13–24, 2015.

[5] E. M. Voorhees, "The trec question answering track," *Natural Language Engineering*, vol. 7, no. 04, pp. 361–378, 2001.

[6] N. A. Smith, M. Heilman, and R. Hwa, "Question generation as a competitive undergraduate course project," in *Proceedings of the NSF Workshop on the Question Generation Shared Task and Evaluation Challenge*, 2008.

[7] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

[8] M. Iyyer, J. Boyd-Graber, L. Claudino, R. Socher, and H. Daumé III, "A neural network for factoid question answering over paragraphs," in *Empirical Methods in Natural Language Processing*, 2014.

[9] J. Berant, A. Chou, R. Frostig, and P. Liang, "Semantic parsing on freebase from question-answer pairs." in *EMNLP*, vol. 2, no. 5, 2013, p. 6.

[10] V. Lopez, C. Unger, P. Cimiano, and E. Motta, "Evaluating question answering over linked data," *Web Semantics Science Services And Agents On The World Wide Web*, vol. 21, pp. 3–13, 2013.

[11] K. M. Hermann, T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Advances in Neural Information Processing Systems (NIPS)*, 2015. [Online]. Available: http://arxiv.org/abs/1506.03340

[12] Y. Yang, W.-t. Yih, and C. Meek, "Wikiqa: A challenge dataset for open-domain question answering," in *Proceedings of EMNLP*. Citeseer, 2015, pp. 2013–2018.

[13] D. Ravichandran and E. Hovy, "Learning surface text patterns for a question answering system," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 41–47.

[14] A. Ittycheriah, M. Franz, W.-J. Zhu, A. Ratnaparkhi, and R. J. Mammone, "Ibm's statistical question answering system." in *TREC*, 2000.

[15] X. Yao and B. Van Durme, "Information extraction over structured data: Question answering with freebase." in *ACL (1)*. Citeseer, 2014, pp. 956–966.

[16] A. Fader, L. S. Zettlemoyer, and O. Etzioni, "Paraphrase-driven learning for open question answering." in *ACL (1)*. Citeseer, 2013, pp. 1608–1618.

[17] M. Iyyer, J. L. Boyd-Graber, L. M. B. Claudino, R. Socher, and H. Daumé III, "A neural network for factoid question answering over paragraphs." in *EMNLP*, 2014, pp. 633–644.

[18] T. Khot, N. Balasubramanian, E. Gribkoff, A. Sabharwal, P. Clark, and O. Etzioni, "Markov logic networks for natural language question answering," *arXiv preprint arXiv:1507.03045*, 2015.

[19] A. B. Abacha and P. Zweigenbaum, "Means: A medical question-answering system combining nlp techniques and semantic web technologies," *Information Processing & Management*, vol. 51, no. 5, pp. 570–594, 2015.

[20] W. Hersh, A. Cohen, L. Ruslen, and P. Roberts, "Trec 2007 genomics track overview."

[21] W. R. Hersh, A. M. Cohen, P. M. Roberts, and H. K. Rekapalli, "Trec 2006 genomics track overview." in *TREC*, 2006.

[22] D. Mollá and J. L. Vicedo, "Question answering in restricted domains: An overview," *Computational Linguistics*, vol. 33, no. 1, pp. 41–61, 2007.

[23] G. Balikas, A. Krithara, I. Partalas, and G. Paliouras, "Bioasq: a challenge on large-scale biomedical semantic indexing and question answering," in *Multimodal Retrieval in the Medical Domain*. Springer, 2015, pp. 26–39.

[24] X. Li and D. Roth, "Learning question classifiers," in *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 2002, pp. 1–7.

[25] R. Morante, M. Krallinger, A. Valencia, and W. Daelemans, "Machine reading of biomedical texts about alzheimers disease 1," 2012.

[26] G. Wiese, D. Weissenborn, and M. Neves, "Neural domain adaptation for biomedical question answering," *arXiv preprint arXiv:1706.03610*, 2017.

TABLE VII. COMPARING BIOMEDICAL QUESTION ANSWERING DATASETS

Dataset	Questions (Train+Test)	Textual Data	Linked Data	Factoid	List	Paragraph/Summary	Yes/No	No. of Entities
TREC 2006 Genomics	28	✓	×	×	×	✓	×	4
TREC 2007 Genomics	36	✓	×	×	×	✓	×	10
QA4MRE about Alzheimer's disease	40	✓	×	✓	✓	×	×	11
QALD-4	25+25	×	✓	✓	×	×	×	3
BioASQ 2013	29+282	✓	✓	✓	✓	✓	✓	∞
BioASQ 2014	310+500	✓	✓	✓	✓	✓	✓	∞
BioASQ 2015	810+500	✓	✓	✓	✓	✓	✓	∞
BioASQ 2016	1307+500	✓	✓	✓	✓	✓	✓	∞
BioASQ 2017	1799+500	✓	✓	✓	✓	✓	✓	∞