RESEARCH

BioASQ and PubAnnotation: Using linked annotations in biomedical question answering

Anastasios Nentidis[1*], Zi Yang[2], Mariana Neves[3], Jin-Dong Kim[4], Anastasia Krithara[1], Georgios Paliouras[1,5] and Ioannis Kakadiaris[5]

[*] Correspondence: tasosnent@iit.demorktios.gr
[1] National Center for Scientific Research "Demokritos", Athens, Greece
Full list of author information is available at the end of the article

Abstract

Motivation: The motivation for this proposal is to extrinsically evaluate the resources available in PubAnnotation and investigate the potential of this repository as an external component of systems with direct real-world biomedical applications, in particular biomedical question answering.

Approach: In this regard, we propose to adjust biomedical question answering systems to take advantage of linked annotations available in PubAnnotation. Those systems can be used to answer biomedical questions from benchmark datasets of the BioASQ challenge, with and without using additional annotations from PubAnnotation. Apart from annotations on relevant biomedical literature, we could also use annotations on benchmark questions through the same repository. Using the BioASQ evaluation infrastructure we can compare the performance of system versions using different resources. Therefore, we can assess the effect of using linked annotations in answering biomedical questions, corresponding to real information needs of biomedical experts.

Keywords: BioASQ; PubAnnotation; biomedical annotations; biomedical question answering

Introduction

BioASQ [1] is a series of challenges on large-scale biomedical semantic indexing and question answering. During the four years of BioASQ running, benchmark biomedical data sets have been developed, for both question answering and semantic indexing, along with a complete infrastructure around them [2]. Regarding the question answering part of the challenge, a team of biomedical experts has developed manually benchmark data sets of biomedical questions with corresponding golden answers and annotations. The PubAnnotation [3] repository provides a uniform and consistent way to store and retrieve linked annotations on biomedical text from different resources or projects. Therefore, we propose a task, in the context of the third Biomedical Linked Annotation Hackathon [4], to investigate if using annotations from this repository can enhance the performance of systems answering biomedical questions from BioASQ benchmark datasets.

Benchmark biomedical questions

Test and evaluation benchmark data sets are available from task B of the BioASQ challenge, consisting of 1799 biomedical questions in English. Those questions are accompanied by gold answers and are also annotated with relevant documents, snippets, concepts and triples, containing the information required to compose their answers.

Question and answer types

Four different types of question are included in the data sets. In particular, yes/no, factoid, list and summary questions. All types of question have ideal answers, which are paragraph-sized summaries, whereas only the first three types have also exact answers, which are short and concise answers (e.g. yes or no, an entity name, or a list of entity names). To begin with, we propose to focus on producing exact answers in the context of this task.

Data set composition

It is expected that annotations will only be available in PubAnnotation for some questions of the original BioASQ data sets. Nevertheless, we propose to use the complete original test sets, as a way of estimating also the completeness of the repository, testing its usefulness in answering a set of questions reflecting real information needs of biomedical experts. In addition, using original test sets, the performance of the systems will be directly comparable to that of challenge participants, using the oracle system of BioASQ [5].

Annotations from PubAnnotation

Systems should be adjusted to retrieve linked annotations, through the PubAnnotation API, and use them to produce answers. Gold relevant documents and snippets, accompanying each question, can be used for accessing annotations. We suggest that the retrieved annotations should be used additionally to annotations from other resources, so that any benefit from them would lead in enhanced performance of the system. In addition to annotations for relevant documents and snippets, it would be very interesting to investigate the potential of using annotations for BioASQ benchmark questions through PubAnnotation. In this regard, the BioMedLAT Corpus [6] will be a valuable resource, consisting of 643 questions from the BioASQ datasets, manually annotated with UMLS semantic types [7].

Annotation types

Different types of annotation are supported in PubAnnotation and are useful for different steps in the pipeline of a question answering system. For example, part-of-speech tags can be used in parsing components of snippets and documents, whereas more informative entities, such as UniProt identifiers, are useful for a concept retrieval component. PubAnnotation also supports “relation annotations” which will be very useful for systems performing reasoning to deduce answers for the questions.

Systems

Interested participants are welcome to adjust any question answering system to contribute to the task. Having results from multiple systems would provide a better understanding of the potential of linked annotations in biomedical question answering, regardless of system and implementation details. However, we suggest that the OAQA Biomedical Question Answering (BioASQ) System [8, 9] is highly suited to this task. This system is already compatible with the benchmark data format and had top performance in exact answer generation for list and factoid questions, in the last two years of the BioASQ challenge [10, 11]. In addition, it has a modular structure which allows the kind of adjustment proposed in this task and it is open source and well documented [12].

Acknowledgements
BioASQ is supported by the National Library of Medicine of the National Institutes of Health and Atypon Systems inc. BioASQ project has also received funding from the European Commission's Seventh Framework Programme.

Author details
[1]National Center for Scientific Research "Demokritos", Athens, Greece. [2]Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. [3]Hasso-Plattner Institute, Potsdam, Brandenburg, Germany. [4]Database Center for Life Science, Research Organization of Information and Systems, Kashiwa, Japan. [5]University of Houston, Texas, USA.

References

1. BioASQ: A Challenge on Large-scale Biomedical Semantic Indexing and Question Answering. http://www.bioasq.org/
2. Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M.R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artieres, T., Ngonga, A., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I., Paliouras, G.: An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics **16**, 138 (2015). doi:10.1186/s12859-015-0564-6
3. PubAnnotation: Make Your Annotation Public, and More Useful! http://www.pubannotation.org/
4. Biomedical Linked Annotation Hackathon (BLAH3). http://blah3.linkedannotation.org
5. BioASQ Participants Area. http://participants-area.bioasq.org/
6. Neves, M., Kraus, M.: Biomedlat corpus: Annotation of the lexical answer type for biomedical questions. In: Open Knowledge Base and Question Answering Workshop at the 26th International Conference on Computational Linguistics (Coling) (2016)
7. The UMLS Semantic Network. https://semanticnetwork.nlm.nih.gov/
8. Yang, Z., Gupta, N., Sun, X., Xu, D., Zhang, C., Nyberg, E.: Learning to answer biomedical factoid and list questions oaqa at bioasq 3b. In: Working Notes for the Conference and Labs of the Evaluation Forum (CLEF), Toulouse, France (2015)
9. Yang, Z., Zhou, Y., Nyberg, E.: Learning to answer biomedical questions: Oaqa at bioasq 4b. ACL 2016, 23 (2016)
10. Balikas, G., Kosmopoulos, A., Krithara, A., Paliouras, G., Kakadiaris, I.: Results of the bioasq tasks of the question answering lab at clef 2015. In: CLEF 2015 (2015)
11. Krithara, A., Nentidis, A., Paliouras, G., Kakadiaris, I.: Results of the 4th edition of bioasq challenge. In: Proceedings of the Fourth BioASQ Workshop at the Conference of the Association for Computational Linguistics, pp. 1–7 (2016)
12. OAQA Biomedical Question Answering (BioASQ) System. https://github.com/oaqa/bioasq