An Adaption of BIOASQ Question Answering dataset for Machine Reading systems by Manual Annotations of Answer Spans

Sanjay Kamath
LIMSI, LRI
Univ. Paris-Sud, CNRS
Université Paris-Saclay
Orsay, France
sanjay@lri.fr

Brigitte Grau
LIMSI, CNRS
ENSIIE
Université Paris-Saclay
Orsay, France
brigitte.grau@limsi.fr

Yue Ma
LRI
Univ. Paris-Sud, CNRS
Université Paris-Saclay
Orsay, France
yue.ma@lri.fr

Abstract

BIOASQ Task B Phase B challenge focuses on extracting answers from snippets for a given question. The dataset provided by the organizers contains answers, but not all their variants. Henceforth a manual annotation was performed to extract all forms of correct answers. This article shows the impact of using all occurrences of correct answers for training on the evaluation scores which are improved significantly.

1 Introduction

BIOASQ[1] challenge is a large-scale biomedical semantic indexing and question answering task (Tsatsaronis et al., 2015) which has been successful for 5 years. The challenge proposes several tasks using Biomedical data. One of the tasks focuses on Biomedical question answering (Task B Phase B - we further refer it as B) where the goal is to extract answers for a given question from relevant snippets.

Several teams have participated actively, and a noticeable aspect is that the results of the task B are much lower compared to open domain QA evaluations, as in SQUAD[2]. Some reasons can be the low dataset size and the format of the answers provided by the organizers. Bioasq provides only certain answer forms in the gold standard data and not all the variants of the answers in the given snippets.

In this paper, we study the influence of enriching the training data by manually annotated variants of gold standard answers on the evaluation performance. We show the impact of the enriched data by experimenting on 5B and 6B training datasets. Our method outperforms the best-performing systems from Bioasq 5B by 7.3% on strict accuracy and 18% on lenient accuracy.

2 Related Work

Several works in the past BIOASQ tasks have used classical question answering pipeline architecture adapted to the biomedical domain. Some use the domain-specific information from UMLS tools such as Metamap (Schulze et al., 2016), along with other NLP tools like Corenlp, LingPipe (Yang et al., 2016). A typical question answering pipeline consists of:

1. Question processing for question type detection and lexical answer type detection.

2. Document retrieval (Task B Phase A)

3. Answer extraction by answer re-ranking on the candidate answers generated in the previous phases, done in a supervised learning manner.

In the open domain, deep learning models are extensively used in machine reading task. Datasets such as MS Marco by (Nguyen et al., 2016), SQUAD by (Rajpurkar et al., 2016) and Wikireading by (Hewlett et al., 2016) have made it easier for deep learning models to perform better on machine reading task. One of the first attempts to use deep learning algorithms for the Bioasq task was reported in BIOASQ 5 by (Wiese et al., 2017b) where the dataset was adapted to be used as a machine reading dataset whose goal is to extract answers from snippets. The authors use a model trained on open domain questions, and perform domain adaptation to biomedical domain using BIOASQ data. Their system got one of the best results whose methods are reported in the section 5.

[1] http://bioasq.org/

[2] https://rajpurkar.github.io/SQuAD-explorer/

Proceedings of the 2018 EMNLP Workshop BioASQ: Large-scale Biomedical Semantic Indexing and Question Answering, pages 72–78
Brussels, Belgium, November 1st, 2018.

3 Evaluation and training data

Bioasq 6 is the sixth challenge and the evaluation measures for Bioasq task B has always been the same. Strict Accuracy, Lenient Accuracy and Mean Reciprocal Rank (MRR) are the 3 evaluation measures used. To compute the scores, the exact match of strings between the predictions and the gold standard answers is used to decide if a system answer is correct. Strict accuracy is the rate of top 1 exact answers. Lenient accuracy is the rate of exact answers in top 5 predictions. MRR is the mean reciprocal rank computed on the top 5 system answers. These measures have been the same since the 1st challenge, although the first four challenges had triples and concepts along with snippets in the data. In the last two challenges, only relevant snippets for questions are released.

Similar evaluations are performed in machine reading tasks like in SQUAD where top 1 accuracy and F1 scores are computed by comparing exact matching strings. One main assumption in machine reading task is that the answer strings are substrings of the snippets, which implies that answers have to be extracted from the snippets.

In Bioasq, the answers are curated by human experts by analyzing the triples, concepts, and snippets (or paragraphs). Thus, the Bioasq dataset and evaluation measures are very similar to that of machine reading task, but the major difference apart from the dataset size are the answers instances provided as gold standard which does not contain all the occurrences, abbreviations, different forms of answers which are present in the snippets.

In (Wiese et al., 2017b), the authors transform Bioasq Phase B as a machine reading task with domain adaptation. Gold standard answer strings and their offsets are automatically searched in the snippets for exact match and treated as answers if only they are found in the snippets, i.e., the answer string must be a substring of the snippet. By doing so the dataset size is reduced to 65% of Bioasq 5 train set which was suitable for adaptation. Other 35% of the questions did not have matching answers in the snippets, because of different variants of answers in the snippets, missing abbreviations, or irrelevant snippets.

This snippet annotation method can result in:

- False positive: an answer mentioned in the snippet which does not answer the question.
- False negative: a snippet answers the question but does not have the exact string compared to the gold standard string.

We found that in Bioasq 6B training dataset for factoid questions, 205 out of 619 questions have false negative answers (33% of the dataset) which may result in some problems:

- Less data for learning;
- The model does not learn to extract all the variants;
- Evaluation is done using such gold standard data which will lower the results even though the model is performing well.

Below are some examples for which the answers returned from a reference system is correct (when evaluated manually) but the automatic evaluation classifies it as incorrect.

> **Q: Which calcium channels does ethosuximide target?**
> **P: ..neuropathic pain is blocked by ethosuximide, known to block T-type calcium channels,..**
>
> **Prediction: T-type calcium**
> **Gold standard: T-type calcium channels**

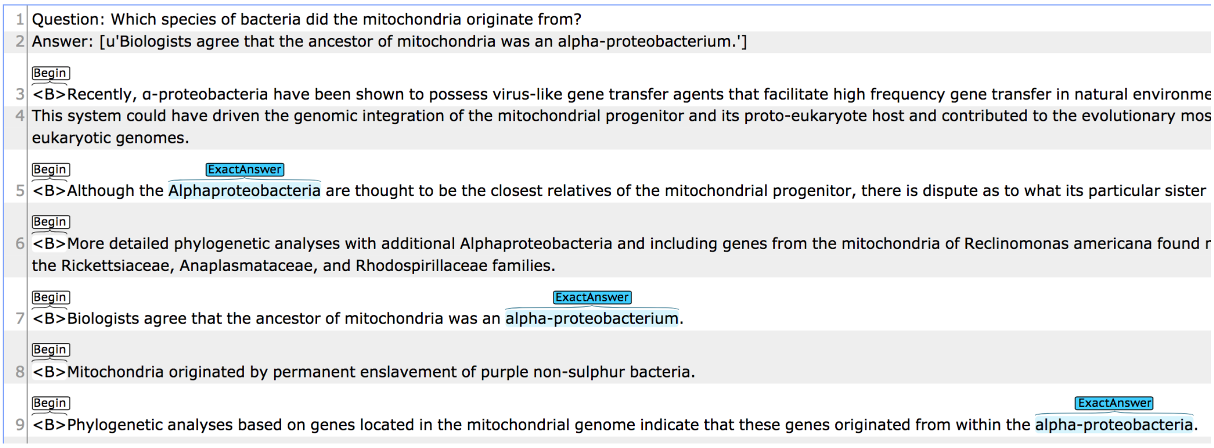
Example 1: Missing keywords

> **Q: Which disease can be treated with Delamanid?**
> **P: In conclusion, delamanid is a useful addition to the treatment options currently available for patients with MDR-TB.**
>
> **Prediction: MDR-TB**
> **Gold standard: tuberculosis**

Example 2: Abbreviations

In example 1, because of a missing word ”channels”, the predicted answer is marked incorrect. In example 2, MDR-TB stands for *Multi-drug-resistant tuberculosis*, which is from a relevant snippet but since the gold standard has only *tuberculosis*, it is marked incorrect. Contextually both are valid answers.

To overcome this problem and enrich the answer space correctly, we manually annotated 618 factoid question-answers pairs from training dataset of 6B task, by annotating the substring of the gold standard answers in the snippets, and adding answers with abbreviations, multi-word answers, synonyms, that are likely correct answers. We explain this in detail in the following section.



```
1 Question: Which species of bacteria did the mitochondria originate from?
2 Answer: [u'Biologists agree that the ancestor of mitochondria was an alpha-proteobacterium.']
3 <B>Recently, α-proteobacteria have been shown to possess virus-like gene transfer agents that facilitate high frequency gene transfer in natural environme
4 This system could have driven the genomic integration of the mitochondrial progenitor and its proto-eukaryote host and contributed to the evolutionary mos
  eukaryotic genomes.
5 <B>Although the Alphaproteobacteria are thought to be the closest relatives of the mitochondrial progenitor, there is dispute as to what its particular sister
6 <B>More detailed phylogenetic analyses with additional Alphaproteobacteria and including genes from the mitochondria of Reclinomonas americana found r
  the Rickettsiaceae, Anaplasmataceae, and Rhodospirillaceae families.
7 <B>Biologists agree that the ancestor of mitochondria was an alpha-proteobacterium.
8 <B>Mitochondria originated by permanent enslavement of purple non-sulphur bacteria.
9 <B>Phylogenetic analyses based on genes located in the mitochondrial genome indicate that these genes originated from within the alpha-proteobacteria.
```

Figure 1: Brat annotation tool

4 Annotations

This section presents the details of the annotations performed manually on the BIOASQ 6B training dataset and presents some statistics.

Our annotations include the following type of answers:

- Exact Answer - Exact match with gold standard (GS) answers, which can also be annotated automatically, and different variants of the answers. For example, the annotation of a single GS answer *"Transcription factor EB (TFEB)"* resulted in 3 annotations, *"Transcription factor EB"*, *"TFEB"*, *"Transcription factor EB (TFEB)"*.

- Lenient Answer - a more general form or a more specific form of an answer. An example is "Telomerase" for "Human telomerase reverse transcriptase".

- Paragraph Answer - The answer matches with gold standard but the snippet alone is not relevant to the question.

We came across several kinds of snippets. A supporting snippet, or answering snippet, is a snippet that contains the answer and enough elements for justifying it. It is a correct answer to the question (snippet starting at line 5 in Figure 1 for example). A snippet that contains the answer without justification towards the question will not be annotated with the answer as correct and is considered as a non-supporting snippet (snippet starting at line 3 in Figure 1). A snippet that does not contain the answer cannot be a supporting snippet, henceforth it is an irrelevant snippet (snippet starting at line 8 in Figure 1).

We use Brat[3] annotation tool by (Stenetorp et al., 2012) shown in Fig. 1 to perform the manual annotations of the snippets with the answer to the question. The annotations done include the answer string along with their character offsets in the snippet. Answers were annotated by 3 people from computer science background and multiple discussions were held to discuss problematic answers which involved looking upon the internet for some medical term meanings.

Annotations were initially done on the Bioasq 5B training set and the additional questions from 5B test sets whose answers are present in the 6B training set were annotated later on 6B data. So the changes done (if any) on 6B training set for previous year questions from 5B set are not considered.

The annotation files are freely available[4] and can be used by researchers who can get the Bioasq dataset.

Some statistics of the dataset are listed in Table 1 for the automatically annotated answers from gold standard data and the fully annotated data with manual annotations. The annotations are done on 618 BIOASQ 6B training dataset questions. Out of 619 factoid questions, 1 question does not have any snippets.

Only 426 questions contain answers from automatic annotation.

"Answers" are the count of answers present in the snippets. *Avg* score represents an average over the total number of questions (i.e. 618). Since in

[3] http://brat.nlplab.org

[4] https://zenodo.org/record/1346193#.W3_WUZMzZQI

	Gold std. annotations		Full annotations	
Count	Avg	Total	Avg	Total
Answers	0.8	500	2.9	1814
Snippets	7.7	3286	8	4965
Questions	-	426	-	618

Table 1: Annotation statistics

gold standard data, only 426 questions have gold answers in snippets, it is normal for the average to fall below 1. It is clear from the table that the full annotated data contain at least 3 times (1814 answers) more the number of candidate answers over the provided gold standard ones (500 answers).

We found that some answers contained the whole snippet as an answer and that 3503 snippets are repeated in the 6B train set. After filtering those repeated snippets we found 3286 different snippets containing exact matching answers extracted automatically from gold standard data and 4965 unique snippets manually annotated with correct answers.

5 Experiments

The goal of our experiments is to study the impact of the data augmentation on training and evaluating a system.

Henceforth, we follow the process of (Wiese et al., 2017b) and use a machine reading model developed by (Chen et al., 2017) that is pre-trained on SQUAD dataset (Rajpurkar et al., 2016) for open domain questions and fine tuned to biomedical questions.

To study the impact on the training process and the evaluations, we train the models using separately the automatically annotated data and the fully manually annotated data. We also evaluate them using both kinds of data separately.

5.1 QA system overview

We present here the adaptation of an existing model named DRQA reader by (Chen et al., 2017) to the biomedical domain as presented in (Kamath et al., 2018).

DRQA reader has three components:

1. Input layer: where the input question words and input passage words are encoded using a pretrained word embedding space.
2. Neural layer: RNN or LSTM networks.
3. Output layer or decoding layer: where the outputs are start and end tokens representing a span of an extracted answer.

The reader model takes as input, the question sentence and the answering snippet and predicts the substring of the snippet that is the answer.

In the input layer, word embeddings are used to encode the words of snippets and questions into vectors, along with textual features such as Part of Speech tags, Named-Entity tokens, Term frequencies of the words in the snippet. The authors use *aligned question embeddings* where an attention score captures the similarity between snippet words and questions words. The neural layer, where the core DNN model is defined, uses different NN architectures to capture semantic similarities between the question/snippet pairs. It use LSTMs to encode the snippets and an RNN to encode the questions. In the output layer, two independent classifiers use a bilinear term to capture the similarity between snippet words and question words and compute the probabilities of each token being start and end of the answer span. We take all possible scores of start and end token predictions and restrict the span between start and end tokens to 15 tokens. We perform an outer product between these scores and consider top 5 spans using an argmax value to get these final predictions.

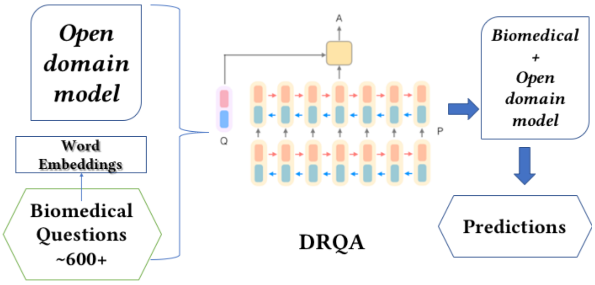


Figure 2: Transfer learning from open domain to biomedical domain

Domain adaptation (also referred to as fine tuning) is performed on the BIOASQ dataset as shown in Figure 2 where the model is pre-trained with SQUAD dataset and fine-tuned with BIOASQ before predicting on test sets. Pre-training is training a model from scratch with randomly initialized weights. Fine-tuning is training on a model with previously trained weights rather than randomly initialized ones. The advantage of pre-

Train set	5B					6B			
Finetune	Gold			Anno.		Gold		Anno.	
Eval	DeepQA	Gold	Anno.	Gold	Anno.	Gold	Anno.	Gold	Anno.
Strict	-	0.2551	**0.2962**	0.1666	**0.3333**	0.2669	**0.3090**	0.2265	**0.3948**
Lenient	-	0.4156	**0.4444**	0.2991	**0.5843**	0.4417	**0.4724**	0.3511	**0.6197**
MRR	0.2620	0.3138	**0.3425**	0.2148	**0.4322**	0.3334	**0.3718**	0.2728	**0.4765**

Table 2: K-fold evaluation on different train sets with *Gold* and *Anno* data. DeepQA scores are presented by (Wiese et al., 2017a)

training with SQUAD dataset is that the DNN model will learn and perform better while trained on a larger training dataset. Since the target dataset is in the biomedical domain, finetuning the previously learned model will have a positive impact on the test set predictions, as shown in (Wiese et al., 2017a).

Several embedding spaces were tested as input vectors (Kamath et al., 2017) and the best performing ones which were the Glove embeddings trained on common crawl data with 840B tokens, were chosen as input to the system. Unknown words were initialized as zero vectors.

As BIOASQ questions have several answering snippets, we treat each question and a snippet as a training sample which might often result in repeated questions with different snippets, i.e. for each training example, there is a question, a unique snippet and the start and end token string offsets of the answer in the snippet. Our model predicts one scored answer per snippet, and the final result is made of the ordered list of answers for the same question. We consider only the top 5 answers.

5.2 Datasets

We perform fine-tuning on two datasets namely

- BIOASQ 5B training set, which contains the 4B training data + the answers of the 4B test data - We term it as 5B.
- BIOASQ 6B training set, which contains the 5B training data + the answers of the 5B test set - We term it as 6B.

We term the automatically annotated training data as *Gold*, and manually annotated training data as *Anno.*

The pre-trained model on open domain QA data is fine-tuned on the above listed Bioasq datasets separately. Evaluation is performed by K-fold cross validation because of the small scale of the data (Table 2), and on the official test sets of Bioasq 5B (Table 3), which were separated from the training data while fine-tuning.

The explanation of scores reported in table 2 and 3 along with the corresponding experiments on the datasets listed above, is as follows. On the data of *Trainset* mentioned in the first row, we fine-tune it with *Finetune* data on the second row - which is *Gold* or *Anno.* version of the answers.

The official evaluation measures[5] using *Gold* or *Anno.* version of the answers are highlighted in the third row. The strict and lenient accuracies along with the MRR are reported.

Gold version of 5B data contains 313 questions and *Gold* version of 6B data contains 428 questions. We consider the remaining questions with no matching answers as incorrectly answered, hence evaluating over all the questions of the datasets (5B - 486 questions, 6B - 618 questions). Annotated 5B data contains 483 questions and 6B data contains 618 questions.

Overall results of 5B test sets presented in Table 3 are evaluated on 150 questions from the test sets of 5B challenge whose gold standard answers are present in 6B challenge train set.

To compare our scores with the ones reported in (Wiese et al., 2017a) and also since the size of the dataset is small, we perform K-Fold (5) evaluations which are reported in Table 2. To compare with previously reported official test scores in Bioasq 5, we train on 5B training set and test on 5B test sets which are reported in Table 3.

6 Results

The results shown in table 2, 3 and 4 highlights the improvements using manually annotated data over the automatically annotated data on the QA performance as well as the evaluations with *Gold*

[5] https://github.com/BioASQ/Evaluation-Measures

Train set	5B					
Finetune			Gold		Anno.	
Eval	(Wiese et al., 2017b)	Lab Zhu, Fudan Univer	Gold	Anno.	Gold	Anno.
Strict	0.3466	0.3533	0.3533	**0.42**	0.3133	**0.4266**
Lenient	0.5066	0.4533	0.54	**0.64**	0.5	**0.6866**
MRR	-	-	0.4256	**0.5042**	0.3884	**0.5258**

Table 3: Overall results calculated on official test sets from 5B task. Scores from (Wiese et al., 2017b) and *Lab Zhu, Fudan Univer* are reported in Bioasq 5.

Train set	5B					
Finetune			Gold		Anno.	
Eval	(Wiese et al., 2017b)	Lab Zhu, Fudan Univer	Gold	Anno.	Gold	Anno.
Batch 1	0.5600	0.4200	0.4733	**0.5733**	0.4933	**0.6066**
Batch 2	0.4086	0.4839	0.4274	**0.5510**	0.3387	**0.5215**
Batch 3	0.4308	0.3846	0.4070	**0.4198**	0.3185	**0.3955**
Batch 4	0.3025	0.2601	0.3595	**0.4474**	0.4444	**0.6196**
Batch 5	0.3924	0.4524	0.4271	**0.4771**	0.3452	**0.5023**

Table 4: MRR results calculated batchwise on 5B official test sets.

and *Anno.* versions of answers.

In Table 2, training on *Gold* and evaluating on *Gold* are the baseline scores. *DeepQA* MRR score is the K-fold evaluation score of MRR reported on 5B train set by (Wiese et al., 2017a).

Comparing the *DeepQA* MRR score with the*Gold* and *Anno.* 5B versions, there is an improvement of at least 17% (*Anno.* training and *Anno.* evaluation) to 8% (*Gold* training and *Anno.* evaluation).

In terms of accuracy, training the model on *Anno.* version and evaluating on *Anno.* version of answers fetch best results by 3.68% and 8.58% on Strict accuracy, 14% and 14.73% on Lenient accuracy in 5B and 6B respectively.

Training on *Anno.* and evaluating on *Gold* has low scores in almost all experiments because of the model which learns on different forms of answers, therefore predicts different forms of answers which are not present in the *Gold* version.

In Table 3, because of a low number of questions in the official test sets ranging from 25 to 35 questions, the scores are computed over all 5B batch test sets by using individual batch results for the number of correct answers from official Bioasq scores and calculating the score over a total number of questions in the 5 batches (5B test sets - 150 questions). The scores by (Wiese et al., 2017b) and *Lab Zhu, Fudan Univer* are the best official results in Bioasq 5. We calculated strict and lenient accuracy as mentioned above and our scores are better than both best official results by 6.67% for strict accuracy and 13.34% lenient accuracy on *Gold* version training, 7.33% for strict accuracy and 18% lenient accuracy on *Anno.* version training.

In Table 4, MRR scores are reported separately for each batch. MRR scores in general have the best scores compared to both (Wiese et al., 2017b) and *Lab Zhu, Fudan Univer* by training on *Anno.* and evaluating on *Anno.* versions.

7 Conclusion and Future Work

We present the importance of using all variants of answers in the snippets for adapting the Bioasq dataset to machine reading task format. We show that the results can be much higher than the officially reported ones if all the variants of the answers are annotated correctly in the training sets. We perform manual annotations to show this impact. Future work would focus on automatic detection of these variants of answers in the snippets.

References

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of ACL 2017*, pages 1870–1879.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew

Kelcey, and David Berthelot. 2016. Wikireading: A novel large-scale language understanding task over wikipedia. *arXiv preprint arXiv:1608.03542*.

Sanjay Kamath, Brigitte Grau, and Yue Ma. 2017. A study of word embeddings for biomedical question answering. In *SIIM'17*.

Sanjay Kamath, Brigitte Grau, and Yue Ma. 2018. Verification of the expected answer type for biomedical question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 1093–1097, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Frederik Schulze, Ricarda Schüler, Tim Draeger, Daniel Dummer, Alexander Ernst, Pedro Flemming, Cindy Perscheid, and Mariana Neves. 2016. Hpi question answering system in bioasq 2016. In *Proceedings of the Fourth BioASQ workshop*, pages 38–44.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138.

Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017a. Neural domain adaptation for biomedical question answering. *arXiv preprint arXiv:1706.03610*.

Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017b. Neural question answering at bioasq 5b. In *BioNLP 2017*, pages 76–79, Vancouver, Canada,. Association for Computational Linguistics.

Zi Yang, Yue Zhou, and Eric Nyberg. 2016. Learning to answer biomedical questions: Oaqa at bioasq 4b. In *Proceedings of the Fourth BioASQ workshop*, pages 23–37.