HAL
open science

Verification of the Expected Answer Type for Biomedical Question Answering

Sanjay Kamath, Brigitte Grau, Yue Ma

To cite this version:

Sanjay Kamath, Brigitte Grau, Yue Ma. Verification of the Expected Answer Type for Biomedical Question Answering. First International Workshop on Hybrid Question Answering with Structured and Unstructured Knowledge (HQA'18), Apr 2018, Lyon, France. pp.1093-1097, 10.1145/3184558.3191542 . hal-01759306

HAL Id: hal-01759306

https://hal.science/hal-01759306

Submitted on 11 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Verification of the Expected Answer Type for Biomedical Question Answering

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

Extractive Question Answering (QA) focuses on extracting precise answers from a given paragraph to questions posed in natural language. Deep learning models are widely used to address this problem and can fetch good results, provided there exists enough data for learning. Such large datasets have been released in open domain, but not in specific domains, such as the medical domain. However, the medical domain has a great amount of resources such as UMLS thesaurus, ontologies such as SNOMED CT, and tools such as Metamap etc that could be useful. In this paper, we apply transfer learning for getting a DNN baseline system on biomedical questions and we study if structured resources can help in selecting the answers based on the recognition of the Expected Answer Type (EAT), which has been proved useful in open domain QA systems. This study relies on different representations for LAT and we study if gold standard answers and answers of our model have some positive impact from the LAT.

KEYWORDS

Question-Answering, Neural Network Model, Expected Answer Type

ACM Reference Format:
Sanjay Kamath, Brigitte Grau, and Yue Ma. 2018. Verification of the Expected Answer Type for Biomedical Question Answering. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3184558.3191542

1 INTRODUCTION

Question Answering (QA) focuses on giving a user precise answers to the questions posed in natural language. In previous evaluation conferences TREC and CLEF, the QA task was, given questions and a large corpus, to provide precise answers extracted from texts with their supporting passage. It involved complex systems, with a pipeline of modules. Since the availability of large training datasets [7, 15, 17], the task has been reformulated as a Machine Reading task: given a question and a paragraph, systems must extract the precise answer in the paragraph. This task is also called Extractive QA. Such a task can be helpful either in open domain, with questions about entities or events, or in specific domains. In this paper, we are interested about medical domain. Even though the lexicon used in these domains are quite different, the types of questions are rather similar and finding information about a gene, a disease or medical events, etc. and information about named entities in open domain have similar structures, and similar approaches can be applied.

One of the large scale datasets for extractive QA in open domain, is SQUAD[1] which was released by [17] consisting of 100,000+ questions posed by crowdworkers on a set of Wikipedia passages. For medical domain, the BIOASQ challenge (Task B) - [18] provide questions with a set of snippets extracted from medical scientific articles. Biomedical questions with their exact answers, relevant text snippets, concepts, articles, summaries were constructed or selected by biomedical experts from around Europe -[13]. An example question with one snippet from BIOASQ data is shown below:

Question:	What is the mode of inheritance of Wilson's disease?
Answer:	autosomal recessive
Snippets:	The overall sex ratio of patients was nearly 1:1, and genetic analysis of 20 families confirmed an autosomal recessive mode of inheritance.

Deep learning models are used widely to address the QA task in open domain and have been proven to be effective. Results of several Deep Neural Network (DNN) models using the SQUAD Dataset can be found on the leaderboard [1]. As it is impossible to create a large scale dataset for biomedical QA without extensive efforts of domain experts, transfer learning can be seen as an alternative approach to use DNN models for small scale biomedical QA as used by [20]. The authors train a deep neural network model based on FASTQA - [19] on open domain data using SQUAD dataset, and then use it to retrain the model on the BioAsq dataset. We use a similar approach for transfer learning with the DRQA model by [3] as it obtains comparable results on open domain QA and its implementation is available[2].

In this paper we are interested to study if adding knowledge belonging to structured resources can help in re-ranking or selecting answers provided by the DNN model. In former QA systems (non deep learning approaches) on text, one of the main criteria for selecting an answer is based on recognizing the Expected Answer Type (EAT) or Lexical Answer Type (LAT) in order to do a matching with candidate answers. It relies on named entity recognition and additional resources, as text corpus and knowledge bases, have been used for improving this verification [4–6]. Biomedical domain has great amount of resources such as UMLS thesaurus, ontologies such

[1]https://rajpurkar.github.io/SQuAD-explorer/
[2]https://github.com/facebookresearch/DrQA

This paper is published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '18 Companion, April 23–27, 2018, Lyon, France
© 2018 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC BY 4.0 License.
ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191542

as SNOMED CT, and tools such as Metamap for annotating texts. Thus, we study if this feature may have a role on getting better answers by studying the relevance of different EAT representations on the corpus provided by [14] by using semantic groups of UMLS and word embeddings of the EAT.

2 RELATED WORK

2.1 QA systems

Former QA systems on text are made of several pipeline modules: question analysis, passage selection, answer selection. Question analysis allows to extract features that are used for selecting passages and extracting the answer. Apart from the content words, these features can be different from a system to another, but they all make use of the Expected Answer Type (EAT) [10]. The EAT is either a named entity type organized in an answer type taxonomy, [11] for open domain or UMLS terminology for bio-medical domain [22], or a word found in the question or a general category as NP (noun phrase) when no information is given about it. Best methods for verifying if a candidate answer matches the EAT, other than NP, involved feature based supervised learning based on the use of different resources, as co-occurrences and presence in structured resources [4–6]. In medical domain, this verification was made using UMLS [2, 22].

Recent QA approaches are based on deep neural network architectures, mainly in the open domain (see results of such models on SQUAD Dataset on the leaderboard [1]). On medical domain Wiese et al. apply domain adaptation for their participation to BIOASQ 2017 and in [19] introduce as supplementary feature the embedding for EAT, defined as the question word or the word close to the question word. However they did not report results that allows to evaluate the impact of EAT.

2.2 Resources

UMLS. (Unified Medical Language System) Metathesaurus, created in 1986, has become an important and a large resource for biomedical science. It provides over 3,100,000 biomedical concepts imported from nearly 200 vocabularies. Each concept is assigned a Concept Unique Identifier (CUI) that uniquely identifies a single meaning. To consistently categorize these huge number of concepts, 133 Semantic Types are defined in UMLS Metathesaurus. In order to further reduce the complexity of the Metathesaurus, these semantic types are divided into 14 groups, called Semantic Groups, as presented in https://semanticnetwork.nlm.nih.gov/download/SemGroups.txt.

Semantic types and semantic groups have been used in various biomedical information system, including categorizing clinical research eligibility criteria [1], learning biomedical ontology [16], and representing clinical questions for medical QA [9].

3 QA SYSTEM OVERVIEW

We present here the adaptation of an existing model named DRQA reader by [3] to the biomedical domain.

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.

In the input layer, word embeddings are used to encode the words of paragraphs and questions into vectors, along with textual features such as Part of Speech tags, Named-Entity tokens, Term frequencies of the words in the paragraph. Authors use *Aligned question embeddings* where an attention score captures the similarity between paragraph words and questions words. Neural layer, where the core DNN model is defined uses different NN architectures to capture semantic similarities between the QA pairs. In the output layer, two independent classifiers use a bilinear term to capture the similarity between paragraph words and question words and compute the probabilities of each token being start and end of the answer span. An argmax value over the unnormalized exponential is calculated on the spans, to get a final prediction.

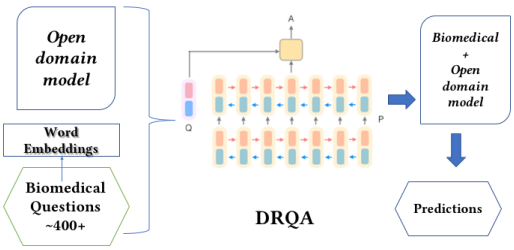

Figure 1: Transfer learning from open domain to biomedical domain

We apply transfer learning as in [20] where the authors train a neural network model based on FASTQA [19] on open domain data using SQUAD dataset, and then use it to retrain the model on the BIOASQ dataset as shown in the Figure 1. Following this model, we first train the DRQA model on SQUAD dataset with its default hyperparameters, and then retrain the model on BIOASQ questions. Several embedding spaces were tested as input vectors [8] and the best performing ones which were the Glove embeddings, were chosen as input to the system.

The BioAsq task is a little different from the SQUAD task. BioAsq provides several snippets that have been considered as relevant by medical experts. Thus, for a same question, the system takes as input each pair with the question and a snippet. In that way, several answers will be predicted for a question from several snippets. We can also notice that some of the snippets does not contain the answer, or the answer is not justified regarding the question, i.e. the snippet contains relevant but non-answerable text extract. Our model predicts one scored answer by snippet, and the final result is made of the ordered list of answers for a same question. We will keep the 5 top answers to study them regarding the EAT representation.

The DNN model does not make use any LAT information or any medical domain related resources. Thus the goal is to study if it can be interesting to add information regarding the LAT, or if the embeddings and the attention model of the model already capture such information.

4 VERIFICATION OF THE EXPECTED ANSWER TYPE

Expected Answer Type or Lexical Answer Type (LAT) helps to identify the type of answers which are to be returned for a question. EAT can be a named entity, number, address, year etc. in open domain, and medical entities such as disease names, genes, drugs, symptoms etc. in biomedical domain. Here is an example below.

Question: What *disease* in Loxapine prominently used for?
Answer: schizophrenia
Expected/Lexical Answer Type: disease
Semantic Group: DISO Disease or Syndrome

In this regard, [14] released a corpus of LAT annotations for BioAsq questions[3] which were manually annotated with LAT words into them and their semantic types from UMLS.

4.1 Material

In our experiments, we consider different representations for the LAT:

- the semantic group the LAT refers to, which can be inferred from semantic types from the UMLS semantic network[4] (SGLAT);
- a word embedding LAT (WELAT).

For computing WELAT when the LAT is made of several words, we compute the average of each word embedding of the LAT. When a word has no embedding, we set its vector to 0. We use Word2Vec skipgram model with 300 dimensions from [12] for computing word embeddings on the biomedical texts of BioAsq 5A task data which consist of 12.8 Million PUBMED articles.

To determine if the recognition of the LAT can be useful for selecting an answer, we study if we can match the LAT given in the annotated corpus (the gold standard LAT for questions (GoldLAT)) with the expected answers (the answers in the gold standard (GoldAns)) and with the answers given by our QA system (PredAns).

First we compute the semantic groups for the GoldLATs by implying the relations between semantic types and semantic groups in the UMLS semantic network[5]. Then we annotate the answers by using MetaMap to obtain the semantic groups of each answers if they exist. We get the correct answers to the questions of the LAT corpus from the Gold Standard data given by BioAsq organizers. We add to these lists the different forms of the answers found in the snippets (short forms, abbreviations etc.). The objective is to perform a realistic automatic evaluation[6].

4.2 Experiments and Results

For the experiments, we consider only the factoid questions from BiomedLat corpus. We split the dataset into train and test set (80% train and 20% test). The statistics reported in the Figure 2 are for the factoid question test set.

[3]https://github.com/mariananeves/BioMedLAT

[4]https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml

[5]https://metamap.nlm.nih.gov/Docs/SemGroups_2013.txt

[6]Note that we did the same annotation of all the training data so that the models built on such datasets should learn from all kinds of forms of answers

We compute cosine similarities between LAT word embeddings in questions and three different answer word embeddings which are detailed below:

- GoldStandard-maxCosine (crossed points): Answer words are Gold standard data annotated with all answer representations that have the maximal cosine similarity with WELAT.
- DRQA-cosine-top1 (triangular points): Answer words are top 1 answers from DRQA output. The similarities of correct (resp. false) answers are plotted above (resp. below) the X-axis.
- DRQA-maxCosine (round points): Answer words are from top 5 answers of DRQA output that have the maximal cosine similarity with WELAT. The similarities of correct (resp. false) answers are plotted above (resp. below) the X-axis.

From Figure 2, we can see that gold standard answers (GoldStandard-maxCosine) show a significant correlation with LAT in terms of word embeddings, although there are 6 questions whose LAT have 0 similarity with WELAT caused by missing word embeddings for the medical domain vocabulary. Another clear observation is that many of top-1 wrong answers from DRQA system have low similarities (less than 0.25), which indicates that we could remove some wrong answers according to this criterion.

Moreover, Figure 2 shows that there are around 50% top-1 answers having zero similarity with question LAT. This could be caused by the out-of-vocabulary problem of word embeddings such as short answers with specific words that have never appeared in the training corpus.

For the round points below the X-axis, they also present an important similarity (around 0.5) correlation with WELAT, which means that by simply selecting the answer with highest similarity as the best answer is not an effective strategy. Indeed, when we used this re-ranking strategy to select one answer from DRQA candidate answers, the strict accuracy with respect to the annotated gold standard decreased from 38% to 33%. Again, the missing word embedding for correct answers has a strong impact on this results.

The observations above show that a fine-grained study of word embeddings is important for medical QA systems.

Table 1: SGLAT associated to answers

Dataset	Answer count
Gold standard data	40/59
DRQA correct top-1 output	18/23
DRQA wrong top-1 output	16/36

To determine the importance of SGLAT in answer words, we studied if the semantic group of the question LAT words are present in the answers. We report this on three datasets, one being the Gold standard questions in BIOMEDLAT corpus and other two being the correct and wrong outputs of DRQA system (top-1).

Table 1 shows the count of matches of SGLAT and answer words. It is clear that many correct answers (gold standard - 40/59) have a matching SGLAT. For DRQA outputs, we compute how many correct and wrong top-1 answers has a matching SGLAT. From the reported findings, there are more correctly answered DRQA outputs (18/23) with matching SGLAT than the wrong ones.

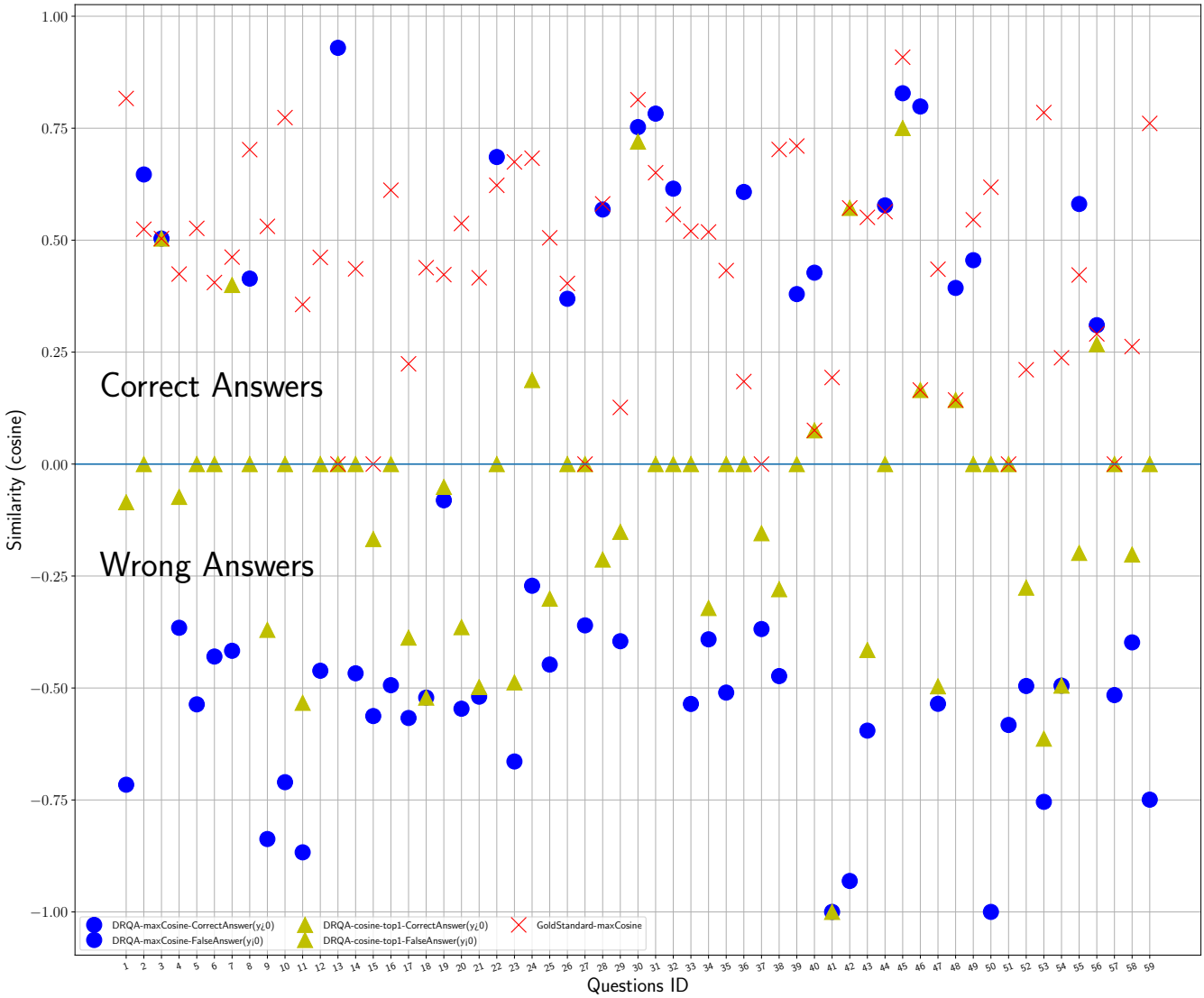

Figure 2: The distribution of answers in three different answering settings for 59 questions: the red crossed points are for gold-standard answers that have the maximal similarity with the question LAT word embedding; So all the crossed points are correct answers. The blue round points above the X-axis are correct answers returned by DRQA system with maximal cosine similarity with WELAT; The round points under the X-axis are false answers found by DRQA system with maximal cosine similarity. The absolute value is the similarity. The green triangles stand for the top-1 results of DRQA system, where the upper parts are correct answers and the low parts are wrong answers.

Using the semantic group annotations from UMLS should have positive impact on the performance of QA systems.

5 CONCLUSION

The expected type of answer of questions has proved to be very useful in former feature-based QA systems as it allows to select correct answers in texts according to their matching type. Nowadays, end-to-end neural network models have been successfully developed for answering questions, in particular in open domain where large datasets were released. These models avoid complex feature engineering. However their adaptation to a specific domain, as medical (bio-medical) domain, by transfer learning shows lower results. Thus we wanted to study if adding some information about LAT could help for improving their results. In this paper, we studied different representations of the LAT, based on structured taxonomy or word embeddings, and showed a correlation with the correct answers. When comparing with the answers provided by our model, we can show that wrong answers might be withdrawn when adding such a criterion and that the computation of word embedding for biomedical terms has to be improved for a neural network based

QA system. In the future, we will study on how we can model this information in our system.

ACKNOWLEDGEMENTS

This work is funded by the ANR project GoAsQ (ANR-15-CE23-0022).

REFERENCES

[1] 2011. Dynamic categorization of clinical research eligibility criteria by hierarchical clustering. *Journal of Biomedical Informatics* 44, 6 (2011), 927 – 935.

[2] Asma Ben Abacha and Pierre Zweigenbaum. 2015. MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. *Information processing & management* 51, 5 (2015), 570–594.

[3] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of ACL 2017*. 1870–1879.

[4] Jennifer Chu-Carroll, James Fan, BK Boguraev, David Carmel, Dafna Sheinwald, and Chris Welty. 2012. Finding needles in the haystack: Search and candidate generation. *IBM Journal of Research and Development* 56, 3.4 (2012), 6–1.

[5] Arnaud Grappy and Brigitte Grau. 2010. Answer type validation in question answering systems. In *RIAO 2010, 9th International Conference, Paris, France, April 28-30, 2010, Proceedings*. 9–15. http://portal.acm.org/citation.cfm?id=1937058&CFID=17354760&CFTOKEN=88565769

[6] Arnaud Grappy, Brigitte Grau, Mathieu-Henri Falco, Anne-Laure Ligozat, Isabelle Robba, and Anne Vilnat. 2011. Selecting answers to questions from Web documents by a robust validation process. In *WI*.

[7] 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* (2016).

[8] Sanjay Kamath, Brigitte Grau, and Yue Ma. 2017. A Study of Word Embeddings for Biomedical Question Answering. In *SIIM'17*.

[9] Tetsuya Kobayashi and Chi-Ren Shyu. 2006. Representing clinical questions by semantic type for better classification. In *AMIA Annual Symposium Proceedings*. 987âĂŞ987.

[10] Oleksandr Kolomiyets and Marie-Francine Moens. 2011. A survey on question answering technology from an information retrieval perspective. *Information Sciences* 181, 24 (2011), 5412–5434.

[11] Xin Li and Dan Roth. 2006. Learning question classifiers: the role of semantic information. *Natural Language Engineering* 12, 3 (2006), 229–249.

[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[13] Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, Georgios Paliouras, and Ioannis Kakadiaris. 2017. Results of the fifth edition of the BioASQ Challenge. In *BioNLP 2017*. 48–57. http://www.aclweb.org/anthology/W17-2306

[14] Mariana Neves and Milena Kraus. 2016. BioMedLAT corpus: Annotation of the lexical answer type for biomedical questions. In *Proceedings of the Open Knowledge Base and Question Answering Workshop (OKBQA 2016)*. 49–58.

[15] 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* (2016).

[16] Alina Petrova, Yue Ma, George Tsatsaronis, Maria Kissa, Felix Distel, Franz Baader, and Michael Schroeder. 2015. Formalizing biomedical concepts from textual definitions. *J. Biomedical Semantics* 6 (2015), 22.

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

[18] 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 (30 Apr 2015), 138. https://doi.org/10.1186/s12859-015-0564-6

[19] Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Making neural qa as simple as possible but not simpler. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. 271–280.

[20] Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017. Neural Domain Adaptation for Biomedical Question Answering. In *Proceedings of CoNLL 2017*. 281–289. https://doi.org/10.18653/v1/K17-1029

[21] Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017. Neural question answering at bioasq 5b. *arXiv preprint arXiv:1706.08568* (2017).

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