Overview of the Medical Question Answering Task at TREC 2017 LiveQA

Asma Ben Abacha[1], Eugene Agichtein[2], Yuval Pinter[3] & Dina Demner-Fushman[1]

(1) U.S. National Library of Medicine, Bethesda, MD
(2) Emory University, Atlanta, GA
(3) Georgia Institute of Technology, Atlanta, GA

Abstract

We present an overview of the medical question answering task organized at the TREC 2017 LiveQA track. The task addresses the automatic answering of consumer health questions received by the U.S. National Library of Medicine. We provided both training question-answer pairs, and test questions with reference answers[1]. All questions were manually annotated with the main entities (foci) and question types. The medical task received eight runs from five participating teams. Different approaches have been applied, including classical answer retrieval based on question analysis and similar question retrieval. In particular, several deep learning approaches were tested, including attentional encoder-decoder networks, long short-term memory networks and convolutional neural networks. The training datasets were both from the open domain and the medical domain. We discuss the obtained results and give some insights for future research in medical question answering.

1 Introduction

The LiveQA track at TREC started in 2015 [2] focusing on answering user questions in real time. The medical QA task was introduced in 2017 based on questions received by the U.S. National Library of Medicine (NLM).

The NLM is the world's largest biomedical library, conducting research, development, and training in biomedical informatics and health information technology. The NLM receives more than 100,000 requests a year, including over 10,000 consumer health questions. The medical task at TREC 2017 LiveQA was organized in the scope of the CHQA project[2] which addresses the classification of customers' requests and the automatic answering of Consumer Health Questions (CHQs).

[1] https://github.com/abachaa/LiveQA_MedicalTask_TREC2017
[2] https://lhncbc.nlm.nih.gov/project/consumer-health-question-answering

CHQs cover a wide range of questions on diseases, drugs, or medical procedures (e.g. Information, Treatment, Comparison, Cause, Usage, Tapering). The question below presents a concrete example of a CHQ looking for *treatments* of "retinitis pigmentosa":

- **Example:**
  Subject: ClinicalTrials.gov - Compliment.
  Message: Hi I have retinitis pigmentosa for 3years. Im suffering from this disease. Please intoduce me any way to treat mg eyes such as stem cell....I am 25 years old and I have only central vision. Please help me. Thank you

Several efforts at the NLM focused on the construction of relevant resources by manually annotating relevant question elements such as the foci and question types [8, 9]. Other research efforts tackled the automatic analysis of consumer health questions [4, 7, 11, 13]. A closely related research area addresses health care-related questions in the context of community-based question answering [10, 15, 17].

2 Task Description

The medical task focuses on providing automatic answers to medical questions. Participants were challenged with retrieving relevant answers to consumer health questions. Two more examples of such questions are presented below. The first CHQ asks about a Problem ("abetalipoproteimemia") and includes more than one subquestion (Diagnosis and Management). The second CHQ includes one subquestion asking about the Ingredients of a Drug (Kapvay).

- **CHQ 1:**
  Subject: abetalipoproteimemia
  Message: hi, I would like to know if there is any support for those suffering with abetalipoproteinemia? I am not diagnosed but have had many test that indicate I am suffering with this, keen to learn how to get it diagnosed and how to manage, many thanks

- **CHQ 2:**
  Subject: ingredients in Kapvay
  Message: Is there any sufites sulfates sulfa in Kapvay? I am allergic.

Question Analysis. One of the main approaches to question answering is extracting the relevant question elements that can lead to correct answers, such as the question focus and type [5]. Other approaches rely on retrieving similar or equivalent questions which were

previously answered [12].

Consumer health questions may contain multiple foci and question types. Users can also describe general and background information such as their medical history before asking their questions, which increases the number of potentially irrelevant medical entities mentioned in the question.

Answer Retrieval. If the question contains more than one subquestion, complete answers should cover all subquestions.

For the medical domain, we suggested the use of trusted medical websites to find relevant answers such as Pubmed abstracts and NIH websites (e.g. `ninds.nih.gov`, `rarediseases.info.nih.gov`, `cancer.gov`). In LiveQA'17, participants were free to use other answer sources such as Quora, Wikipedia or medical websites where doctors answer online questions (e.g. `icliniq.com`).

3 Training Datasets

We provided two training sets with 634 pairs of medical questions and answers. We also provided additional annotations for the Question Focus and the Question Type used to define each subquestion. Training questions cover four categories of foci (Disease, Drug, Treatment and Exam) and 23 question types (e.g. Treatment, Cause, Indication, dosage).

The first training dataset consists of 388 (sub)question-answer pairs corresponding to 200 NLM questions. Figure 1 presents an example from this training dataset.

Each question is divided into one or more subquestion(s). Each subquestion has one or more answer(s). QA pairs were constructed from FAQs on trusted websites of the U.S. National Institutes of Health (NIH). Candidate question-answer pairs were retrieved using automatic matching between the CHQs and the FAQs based on the focus and the question type. The QA pairs retained for training are the manually validated pairs from the candidate set.

The second training dataset consists of 246 question-answer pairs corresponding to 246 NLM questions. Answers were retrieved manually by librarians using PubMed and web search engines.

4 Test Dataset

Test Questions. The test set consists of 104 NLM questions. The subquestion, focus and type annotations were not provided to the participants. For each medical question, participants were tasked to retrieve a correct answer for each subquestion. If the question includes more than one subquestion, answers should be ranked according to the order of the subquestions.

We selected the test questions to cover a wide range of question types (26) and have a slightly different distribution than the training questions in order to evaluate the scalability of the proposed systems. Section 4.1 describes in more details the question types and the

```
−<NLM-QUESTION questionid="Q10" fRef="1-135752923">
  <SUBJECT/>
 −<MESSAGE>
    Would appreciate any good info on Lewy Body Dementia, we need to get people aware of this dreadful disease, all they talk about is alzheimers.
    Thank you
  </MESSAGE>
 −<SUB-QUESTIONS>
  −<SUB-QUESTION subqid="Q10-S1">
   −<ANNOTATIONS>
      <FOCUS>lewy body dementia</FOCUS>
      <TYPE>information</TYPE>
    </ANNOTATIONS>
   −<ANSWERS>
    −<ANSWER answerid="Q10-S1-A1" pairid="17">
        Summary Lewy body disease is one of the most common causes of dementia in the elderly. Dementia is the loss of mental functions severe enough to affect normal activities and relationships. Lewy body disease happens when abnormal structures, called Lewy bodies, build up in areas of the brain. The disease may cause a wide range of symptoms, including - Changes in alertness and attention - Hallucinations - Problems with movement and posture - Muscle stiffness - Confusion - Loss of memory Lewy body disease can be hard to diagnose, because Parkinson's disease and Alzheimer's disease cause similar symptoms. Scientists think that Lewy body disease might be related to these diseases, or that they sometimes happen together. Lewy body disease usually begins between the ages of 50 and 85. The disease gets worse over time. There is no cure. Treatment focuses on drugs to help symptoms. NIH: National Institute of Neurological Disorders and Stroke
      </ANSWER>
    −<ANSWER answerid="Q10-S1-A2" pairid="18">
        Lewy body dementia is one of the most common forms of progressive dementia. People affected by this condition may experience a variety of symptoms such as changes in alertness and attention; hallucinations; problems with movement and posture; muscle stiffness; confusion; and/or memory loss. Although the exact cause of Lewy body dementia is poorly understood, symptoms are thought to result when clumps of a protein called alpha-synuclein ("Lewy bodies") accumulate in the brain. Lewy body dementia usually occurs sporadically in people with no family history of the condition. Rarely, more than one family member may be affected. There is currently no cure for Lewy body dementia; however, medications may be available to help manage the associated symptoms.
      </ANSWER>
    </ANSWERS>
  </SUB-QUESTION>
 </SUB-QUESTIONS>
</NLM-QUESTION>
```


Figure 1: Annotated example from the first training dataset.

foci categories associated with the test set.

Reference Answers. For each test question, we manually collected one or more reference answer(s) from trusted sources such as NIH websites. NIST assessors created question paraphrases/interpretations after reading both the original questions and the reference answers. They used the paraphrases with the reference answers to judge the participants' answers. Below is an example of a consumer health question with the associated reference answer:

- **Question Subject:** Can cancer spread through blood contact
- **Question Message:** Sir, after giving an insulin injection to my uncle who is a cancer patient the needle accidentally pined my finger. Is there a problem for me? Plz reply.
- **Reference Answer:** A healthy person cannot "catch" cancer from someone who has it. There is no evidence that close contact or things like sex, kissing, touching, sharing meals, or breathing the same air can spread cancer from one person to another. Cancer cells from one person are generally unable to live in the body of another healthy person. A healthy person's immune system recognizes foreign cells and destroys them, including cancer cells from another person.
- **Answer URL:** *https://www.cancer.org/cancer/cancer-basics/is-cancer-contagious.html*

4.1 Additional Annotations: Foci, Question Types and Keywords

We provided additional annotations (Foci/Question Types/Keywords) after the challenge, for future efforts and evaluations. The examples below present some of the provided annotations: Foci are highlighted in blue, question types and their triggers in red and keywords in green:

- **Consumer Health Question:**
  Subject: Testing for EDS.
  Message: I would like to know if you can point me in the direction of a laboratory in Southern California, Specifically San Bernardino County or LA County or even Riverside County that does genetic testing for EDS or Osteogenesis Imperfecta and do you know if the two diseases are similiar in symptoms? Thank you for you help and time.
- **Provided Annotations:**
  Q-Focus fid="F1" fcategory="Problem": EDS
  Q-Focus fid="F2" fcategory="Problem": Osteogenesis Imperfecta
  Q-Type tid="T1" hasFocus="F1,F2": COMPARISON
  Q-Type tid="T2" hasFocus="F1": DIAGNOSIS
  Q-Type tid="T3" hasFocus="F1" hasKeyword="K1": PERSON_ORGANIZATION
  Q-Type tid="T4" hasFocus="F2": DIAGNOSIS
  Q-Type tid="T5" hasFocus="F2" hasKeyword="K1": PERSON_ORGANIZATION
  Q-Keyword kid="K1" kcategory="GeographicLocation": Southern California

Annotating the test questions also allowed us to provide more detailed statistics about the test set. Figures 2, 3 and 4 present the types of test questions, as well as the categories

associated with the foci and keywords. These categories and types can help improving the scalability of question analysis methods and the coverage of answer resources used for the medical domain. Figure 5 presents an example from the annotated test dataset.

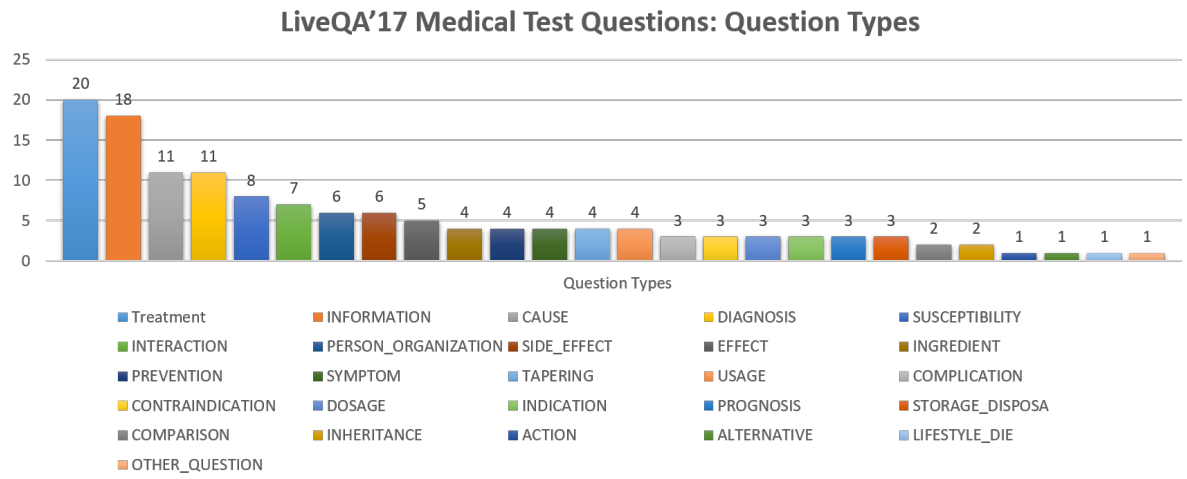

Figure 2: Questions types covered by the medical test questions.

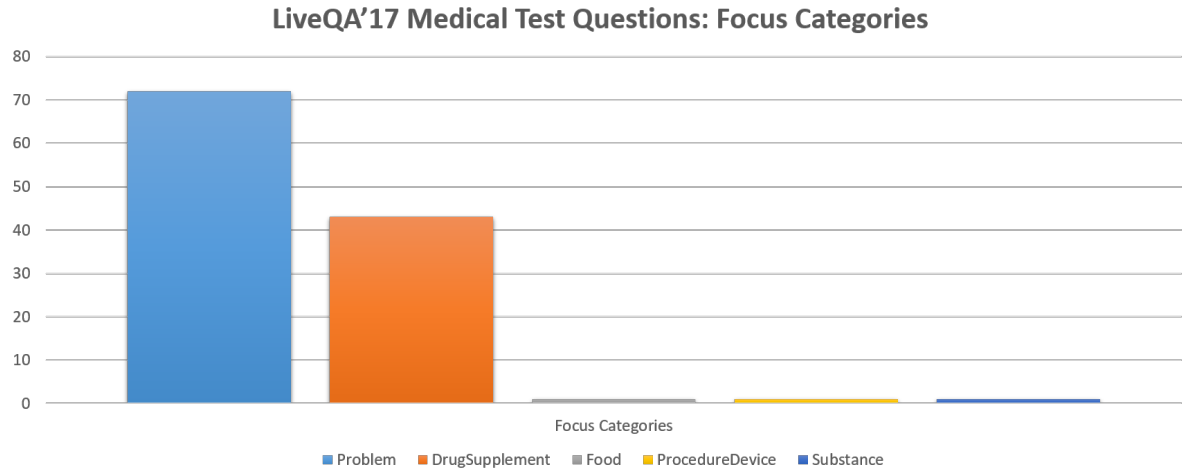

Figure 3: Categories associated with the foci of the medical test questions.

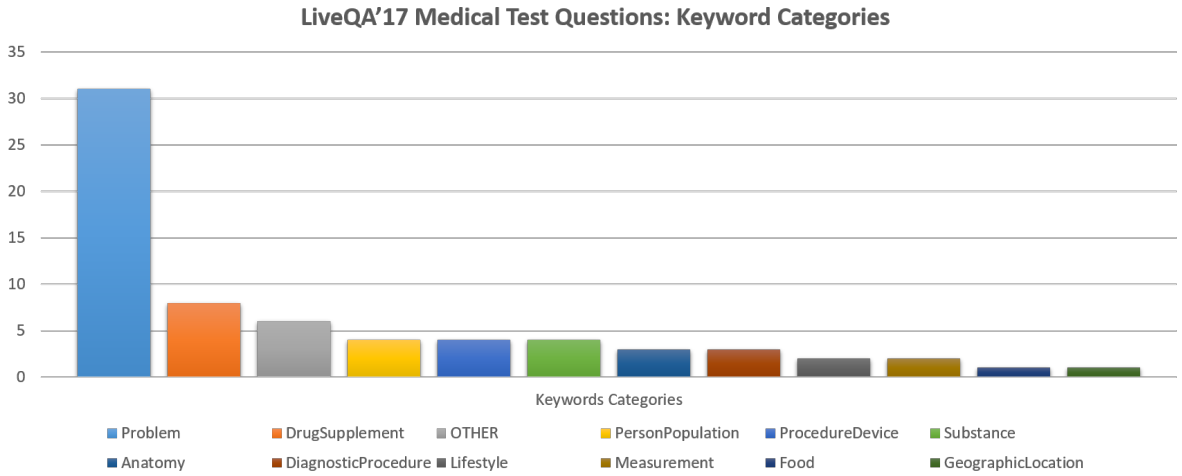

Figure 4: Categories associated with the keywords of the medical test questions.

```
−<NLM-QUESTION qid="TQ19">
  −<Original-Question qfile="1-137067367.xml.txt">
     <SUBJECT>Sevoflurane</SUBJECT>
    −<MESSAGE>
       I work in a hospital, and a question recently came up regarding the stability of Sevoflurane once it has been opened. Does Sevoflurane expire
       within a particular timeframe or is the product still effective until the expiration date listed on the bottle?
     </MESSAGE>
   </Original-Question>
  −<NIST-PARAPHRASE>
     What is the stability, effectiveness and toxicity of sevoflurane once the product container has been opened?
   </NIST-PARAPHRASE>
  −<ANNOTATIONS>
     <FOCUS fid="F1" fcategory="DrugSupplement">Sevoflurane</FOCUS>
     <TYPE tid="T1" hasFocus="F1">USAGE</TYPE>
   </ANNOTATIONS>
  −<ReferenceAnswers>
    −<ReferenceAnswer aid="TQ19A1">
      −<ANSWER>
         We prepared a 20% sevoflurane lipid emulsion using caprylic triglyceride (i.e., medium-chain triglyceride). In rats, this emulsion was an effective
         anesthetic and was not associated with adverse events. The emulsion was stable after consecutive evaluation for 365 days and for 180 minutes
         after the vial was opened.
       </ANSWER>
       <AnswerURL> https://www.ncbi.nlm.nih.gov/pubmed/26716717 </AnswerURL>
      −<COMMENT>
         provides information on stability after opening the vial
       </COMMENT>
     </ReferenceAnswer>
    −<ReferenceAnswer aid="TQ19A2">
      −<ANSWER>
         Sevoflurane is stable when stored under normal room lighting conditions. No discernible degradation of sevoflurane occurs in the presence of
         strong acids or heat. Sevoflurane is not corrosive to stainless steel, brass, aluminum nickel-plated brass, chrome-plated brass or copper
         beryllium alloy.
       </ANSWER>
       <AnswerURL>https://www.medicines.org.uk/emc/medicine/49</AnswerURL>
      −<COMMENT>
         Provides information about stability of the substance in general. This information is also relevant.
       </COMMENT>
     </ReferenceAnswer>
   </ReferenceAnswers>
</NLM-QUESTION>
```

Figure 5: Annotated example from the test dataset.

5 Submissions and Results

5.1 Submissions

Five teams participated in the medical QA task with eight runs in total. Table 1 presents the participating teams and the submitted runs.

Team	Country	Run
Carnegie Mellon University CMU-LiveMedQA	USA	CMU-LiveMedQA
Carnegie Mellon University CMU-OAQA	USA	CMU-OAQA
East China Normal University ECNU	China	ECNU
East China Normal University ECNU-ICA	China	ECNU_ICA_1 ECNU_ICA_2
Philips Research North America PRNA	USA	prna-r1 prna-run2 prna-run3

Table 1: LiveQA 2017 Medical Task: Participating teams and submitted runs

5.2 Results

We use the same scoring scheme as the main TREC LiveQA challenge [1, 2]:

- avgScore [0-3 range]: the average score over all questions, transferring 1-4 level grades to 0-3 scores. This is the main score used to rank LiveQA runs.
- succ@i+: the number of questions with score i or above (i∈2,4) divided by the total number of questions.
- prec@i+: the number of questions with score i or above (i∈2,4) divided by number of questions answered by the system.

The results presented in this section use the number of questions which were answered by all systems (102) as the total number of questions, instead of the original number of test questions (104). Table 2 presents the Average Score and Success. CMU-OAQA achieved the best Average Score of 0.637. Table 3 presents the Precision results.

6 Discussion

The LiveQA track has been running since 2015 at TREC. This year, the medical QA task was introduced focusing on consumer health questions. The proposed test questions cover a wide range of question types and have a slightly different distribution than the training questions to allow evaluating the performance and scalability of the proposed systems.

Participant	AvgScore [0-3]	S@2	S@3	S@4
CMU-OAQA	**0.637**	0.392	0.265	0.098
prna-r1	0.490	0.265	0.157	0.069
prna-run2	0.441	0.275	0.137	0.059
prna-run3	0.431	0.284	0.147	0.059
ECNU_ICA_2	0.402	0.216	0.127	0.059
CMU-LiveMedQA	0.353	0.216	0.137	0
ECNU_ICA_1	0.255	0.225	0.147	0.029
ECNU	0.137	0.216	0.088	0.01

Table 2: Official Results of the Medical QA Task: Average Score and Success

Participant	P@2	P@3	P@4
CMU-OAQA	0.404	0.273	0.101
prna-r1	0.429	0.254	0.111
prna-run2	0.394	0.197	0.085
prna-run3	0.397	0.205	0.205
ECNU_ICA_2	0.268	0.159	0.073
CMU-LiveMedQA	0.218	0.139	0
ECNU_ICA_1	0.228	0.149	0.03
ECNU	0.216	0.088	0.01

Table 3: Official Results of the Medical QA Task: Precision

The CMU-OAQA system [14] achieved the best performance of 0.637 on the medical task. They used an attentional encoder-decoder model for paraphrase identification and answer ranking. Quora question-similarity dataset was used for training.

The PRNA system [6] achieved the second best performance in the medical task with 0.49 avgScore (prna-r1). They used Wikipedia as the first answer source and Yahoo and Google searches as secondary answer sources. Each medical question was decomposed into several subquestions. Then wikipages associated with the question focus were used as candidate answer sources. To extract the answer from the selected text passage, a bi-directional attention model trained on the SQUAD dataset was used.

Another technique was used by ECNU-ICA team [3] based on learning question similarity via two long short-term memory (LSTM) networks applied to obtain the semantic representations of the questions. To construct a collection of similar question pairs, they searched community question answering sites such as Yahoo! and Answers.com.

The CMU-LiveMedQA team [16] designed a specific system for the medical task. Using only the provided training datasets and the assumption that each question contains only one focus, the CMU-LiveMedQA system obtained an avgScore of 0.353. They used a convolutional neural network (CNN) model to classify a question into a restricted set of 10 question types and crawled "relevant" online web pages to find the answers.

Although the number of submitted runs (8) is limited, the tested methods and their results can give some insights and directions. For instance, Open-domain datasets (e.g. Quora, Wikipedia) were helpful for the medical domain. The best system on the medical task (CMU-OAQA) with 0.637 avgScore used only the Quora training dataset. It is also relevant to note that the same system and training data obtained a score of 1.139 on the LiveQA open-domain task [1]. These two results support the relevance of similar question matching for the end-to-end QA task. The gap in performance between the open domain and the medical domain can be explained in part by the discrepancies between the medical test questions and the open-domain questions used for training.

In contrast, the ECNU-ICA system achieved the best performance of 1.895 in the open-domain task but an average score of only 0.402 in the medical task (ECNU_ICA_2). As the ECNU-ICA approach also relied on a neural network for question matching, this result shows that training attention-based decoder-encoder networks on the Quora dataset generalized a lot better to the medical domain than training LSTMs on similar questions from Yahoo! and Answers.com.

More generally, the current gap in performance between the open-domain task and the medical task supports the need for larger medical datasets to support deep learning approaches in dealing with the linguistic complexity of consumer health questions and the challenge of finding correct and complete answers.

7 Conclusion

We described the medical QA task organized at the TREC 2017 LiveQA Track. Two training datasets with pairs of NLM questions and relevant answers were provided. We have also provided reference answers as well as additional annotations of the test questions. All datasets were publicly released to support research efforts in medical question answering[3]. The task attracted five teams that submitted eight runs. Different approaches relying on attention based decoders, LSTMs and CNNs were tested and compared. The obtained results highlight the difficulty of the medical QA task in comparison with open-domain QA, and provide pointers for future research and development.

Acknowledgements

We would like to thank Sonya Shooshan for her help with the reference answers and NIST Assessors for judging participants' answers. We are also thankful to Ellen Voorhees as well as all organizers and participants for their valuable efforts and support.

[3] `https://github.com/abachaa/LiveQA_MedicalTask_TREC2017`

References

[1] Agichtein, E., Ben Abacha, A., Harman, D., Nyberg, E., and Pinter, Y. Overview of the TREC 2017 liveqa track. In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA* (2017).

[2] Agichtein, E., Carmel, D., Pelleg, D., Pinter, Y., and Harman, D. Overview of the TREC 2015 liveqa track. In *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA* (2015).

[3] An, W., Chen, Q., Tao, W., Zhang, J., Yu, J., Yang, Y., Hu, Q., He, L., and Li, B. Ecnu at 2017 liveqa track: Learning question similarity with adapted long short-term memory networks. In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA* (2017).

[4] Ben Abacha, A., and Demner-Fushman, D. Recognizing question entailment for medical question answering. In *AMIA 2016, American Medical Informatics Association Annual Symposium, Chicago, IL, USA, November* (2016).

[5] Ben Abacha, A., and Zweigenbaum, P. MEANS: A medical question-answering system combining NLP techniques and semantic web technologies. *Information Processing and Management Journal 51*, 5 (2015), 570–594.

[6] Datla, V., Arora, T., Liu, J., Adduru, V., Hasan, S. A., Lee, K., Qadir, A., Ling, Y., Prakash, A., and Farri, O. Prna at the TREC 2017 liveqa track. In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA* (2017).

[7] Guo, H., Na, X., Hou, L., and Li, J. Classifying Chinese Questions Related to Health Care Posted by Consumers Via the Internet. *JMIR medical informatics 19*, 6 (2017).

[8] Kilicoglu, H., Ben Abacha, A., Mrabet, Y., Roberts, K., Rodriguez, L., Shooshan, S. E., and Demner-Fushman, D. Annotating named entities in consumer health questions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia* (2016).

[9] Kilicoglu, H., Ben Abacha, A., Mrabet, Y., Shooshan, S. E., Rodriguez, L., Masterton, K., and Demner-Fushman, D. Semantic annotation of consumer health questions. *BMC Bioinformatics 19*, 1 (2018), 34:1–34:28.

[10] Liu, T., Zhang, W., Cao, L., and Zhang, Y. Question Popularity Analysis and Prediction in Community Question Answering Services. *PLoS One 9*, 5 (2014).

[11] Mrabet, Y., Kilicoglu, H., Roberts, K., and Demner-Fushman, D. Combining open-domain and biomedical knowledge for topic recognition in consumer health

questions. In *AMIA 2016, American Medical Informatics Association Annual Symposium, Chicago, IL, USA, November* (2016).

[12] Nakov, P., Màrquez, L., Moschitti, A., Magdy, W., Mubarak, H., Freihat, A. A., Glass, J., and Randeree, B. Semeval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016* (2016).

[13] Roberts, K., Kilicoglu, H., Fiszman, M., and Demner-Fushman, D. Automatically classifying question types for consumer health questions. In *AMIA 2014, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 15-19, 2014* (2014).

[14] Wang, D., and Nyberg, E. Cmu oaqa at trec 2017 liveqa: A neural dual entailment approach for question paraphrase identification. In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA* (2017).

[15] Wongchaisuwat, P., Klabjan, D., and Jonnalagadda, S. R. A Semi-Supervised Learning Approach to Enhance Health Care Community-Based Question Answering: A Case Study in Alcoholism. *JMIR medical informatics 4*, 3 (2016).

[16] Yang, Y., Yu, J., Hu, Y., Xu, X., and Nyberg, E. Cmu livemedqa at trec 2017 liveqa: A consumer health question answering system. In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA* (2017).

[17] Zhang, W., Liu, T., Yang, Y., Cao, L., Zhang, Y., and Ji, R. A Topic Clustering Approach to Finding Similar Questions from Large Question and Answer Archives. *PLoS One 9*, 3 (2014).