BMC Bioinformatics

RESEARCH ARTICLE Open Access

A question-entailment approach to question answering

Asma Ben Abacha* and Dina Demner-Fushman

Abstract

Background: One of the challenges in large-scale information retrieval (IR) is developing fine-grained and domain-specific methods to answer natural language questions. Despite the availability of numerous sources and datasets for answer retrieval, Question Answering (QA) remains a challenging problem due to the difficulty of the question understanding and answer extraction tasks. One of the promising tracks investigated in QA is mapping new questions to formerly answered questions that are "similar".

Results: We propose a novel QA approach based on Recognizing Question Entailment (RQE) and we describe the QA system and resources that we built and evaluated on real medical questions. First, we compare logistic regression and deep learning methods for RQE using different kinds of datasets including textual inference, question similarity, and entailment in both the open and clinical domains. Second, we combine IR models with the best RQE method to select entailed questions and rank the retrieved answers. To study the end-to-end QA approach, we built the MedQuAD collection of 47,457 question-answer pairs from trusted medical sources which we introduce and share in the scope of this paper. Following the evaluation process used in TREC 2017 LiveQA, we find that our approach exceeds the best results of the medical task with a 29.8% increase over the best official score.

Conclusions: The evaluation results support the relevance of question entailment for QA and highlight the effectiveness of combining IR and RQE for future QA efforts. Our findings also show that relying on a restricted set of reliable answer sources can bring a substantial improvement in medical QA.

Keywords: Question Answering, Recognizing Question Entailment, Machine Learning, Deep Learning, Information Retrieval, Consumer Health Questions, Medical Question-Answer Dataset

Background

With the availability of rich data on users' locations, profiles, and search histories, personalization has become the leading trend in large-scale information retrieval. However, efficiency through personalization is not yet the most suitable model when tackling domain-specific searches. This is due to several factors, such as the lexical and semantic challenges of domain-specific data that often include advanced argumentation and complex contextual information, the higher sparseness of relevant information sources, and the more pronounced lack of similarities between users' searches.

A recent study on expert search strategies among healthcare information professionals [1] showed that, for a given search task, they spend an average of 60 min per collection or database, 3 min to examine the relevance of each document, and 4 h of total search time. When written in steps, their search strategy spans over 15 lines and can reach up to 105 lines.

With the abundance of information sources in the medical domain, consumers are increasingly faced with a similar challenge, one that needs dedicated solutions that can adapt to the heterogeneity and specifics of health-related information.

Dedicated Question Answering (QA) systems are one of the viable solutions to this problem as they are designed to understand natural language questions without relying on external information about the users.

In the context of QA, the goal of Recognizing Question Entailment (RQE) is to retrieve answers to a *premise question* (*PQ*) by retrieving inferred or entailed questions, called *hypothesis questions* (*HQ*) that already have

*Correspondence: asma.benabacha@nih.gov
[1]Lister Hill Center, U.S. National Library of Medicine, U.S. National Institutes of Health, Bethesda, MD, USA

associated answers. Therefore, we define the entailment relation between two questions as: a question *A* **entails** a question *B* if every answer to *B* is also a **correct answer** to *A* [2].

In contrast with end-to-end QA approaches like deep learning models for QA, RQE-based QA is a multi-step approach that tackles the challenging issues of question understanding and answer extraction in a unique way. RQE is also particularly relevant due to the increasing numbers of similar questions posted online [3]. In addition to being used to find relevant answers, RQE resources can also be used in training models to recognize inference relations and similarity between questions.

Question similarity has recently attracted international challenges [4, 5] and research efforts proposing a wide range of approaches, including logistic regression, Recurrent Neural Networks (RNNs), Long Short Term Memory cells (LSTMs), and Convolutional Neural Networks (CNNs) [2, 6–8].

In this paper, we study question entailment in the medical domain and the effectiveness of the end-to-end RQE-based QA approach by evaluating the relevance of the retrieved answers. Although entailment was attempted in QA before [9–11], as far as we know, we are the first to introduce and evaluate a full medical question answering approach based on question entailment for free-text questions. Our contributions are:

1. A study of logistic regression and deep learning models applied to RQE using different kinds of datasets, including textual inference, question similarity and entailment in both the open and clinical domains.
2. A collection of 47,457 medical question-answer pairs with additional annotations, constructed from trusted sources such as NIH websites. We make this resource publicly available[1].
3. A new QA approach based on question entailment. Our approach uses IR models to retrieve question candidates and the RQE model to identify entailed questions and return their answers.
4. An evaluation of the RQE-based QA system on TREC 2017 LiveQA medical questions [12]. The results showed that our approach exceeds the best official score on the medical task using only the collection of 47K QA pairs as answers source.

We define below the RQE task and describe related work at the intersection of question answering, question similarity and textual inference.

Task Definition

The definition of Recognizing Question Entailment (RQE) can have a significant impact on QA results. In related work, the meaning associated with Natural Language Inference (NLI) varies among different tasks and events. For instance, Recognizing Textual Entailment (RTE) was addressed by the PASCAL challenge [13], where the entailment relation has been assessed manually by human judges who selected relevant sentences "entailing" a set of hypotheses from a list of documents returned by different Information Retrieval (IR) methods. In another definition, the Stanford Natural Language Inference corpus SNLI [14], used three classification labels for the relations between two sentences: entailment, neutral and contradiction. For the entailment label, the annotators who built the corpus were presented with an image and asked to write a caption "that is a definitely a true description of the photo". For the neutral label, they were asked to provide a caption "that might be a true description of the photo". They were asked for a caption that "is definitely a false description of the photo" for the contradiction label.

More recently, the multiNLI corpus [15] was shared in the scope of the RepEval 2017 shared task[2] [16]. To build the corpus, annotators were presented with a premise text and asked to write three sentences. One novel sentence, which is "necessarily true or appropriate in the same situations as the premise," for the entailment label, a sentence, which is "necessarily false or inappropriate whenever the premise is true," for the contradiction label, and a last sentence, "where neither condition applies," for the neutral label.

Whereas these NLI definitions might be suitable for the broad topic of text understanding, their relation to practical information retrieval or question answering systems is not straightforward.

In contrast, RQE needs tailoring to the question answering task. For instance, if the premise question is "looking for cold medications for a 30 yo woman", a RQE approach should be able to consider the more general (less restricted) question "looking for cold medications" as relevant, since its answers are relevant to the initial question. The entailment relation we are seeking in the QA context should include relevant and meaningful relaxations of contextual and semantic constraints (cf. "Definition" section).

Related Work on Question Answering

Classical QA systems face two main challenges related to question analysis and answer extraction. Several QA approaches were proposed in the literature for the open domain [17, 18] and the medical domain [19–21]. A variety of methods were developed for question analysis,

[1] https://github.com/abachaa/MedQuAD

[2] https://repeval2017.github.io/shared

focus (topic) recognition, and question type identification [22–25]. Similarly, many different approaches tackled document or passage retrieval and answer selection and (re)ranking [26–30].

An alternative approach consists in finding similar questions or frequently asked questions (FAQs) that are already answered [31, 32]. One of the earliest question answering systems based on finding similar questions and re-using the existing answers was FAQ FINDER [33]. Another system that complements the existing Q&A services of NetWellness[3] is SimQ [3], which allows retrieval of similar web-based consumer health questions. SimQ uses syntactic and semantic features to compute similarity between questions, and UMLS [34] as a standardized semantic knowledge source. The system achieves 72.2% precision, 78.0% recall and 75.0% F-score on NetWellness questions. However, the method was evaluated only on one question similarity dataset, and the retrieved answers were not evaluated.

The aim of the medical task at TREC 2017 LiveQA was to develop techniques for answering complex questions such as consumer health questions, as well as to identify reliable answer sources that can comply with the sensitivity of medical information retrieval in terms of its impact on public health.

The CMU-OAQA system [35] achieved the best performance of 0.637 average score on the medical task by using an attentional encoder-decoder model for paraphrase identification and answer ranking. The Quora question-similarity dataset was used for training. The PRNA system [36] achieved the second-best performance in the medical task with 0.49 average score using Wikipedia as the first answer source and Yahoo and Google searches as secondary answer sources. Each medical question was decomposed into several subquestions. To extract the answer from the selected text passage, a bi-directional attention model trained on the SQuAD dataset [37] was used.

Deep neural network models have been pushing the limits of performance achieved in QA related tasks using large training datasets. The results obtained by CMU-OAQA and PRNA showed that large open-domain datasets can be beneficial for the medical domain. Other studies also highlighted the same finding [38]. However, the best system (CMU-OAQA) relying on the same training data obtained a score of 1.139 in the LiveQA open-domain task (vs. 0.637 in the medical task).

While this gap in performance can be explained in part by the discrepancies between the medical test questions and the open-domain questions, it also highlights 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.

Another technique was used by ECNU-ICA team [39] 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. 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. 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 better to the medical domain than training LSTMs on similar questions from Yahoo! and Answers.com.

The CMU-LiveMedQA team [21] 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 average score 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. However, the results were lower than those achieved by the systems relying on finding similar answered questions. These results support the relevance of similar question matching for the end-to-end QA task as a new way of approaching QA instead of the classical QA approaches based on Question Analysis and Answer Retrieval.

Related Work on Question Similarity and Entailment

Several efforts focused on recognizing similar questions. Jeon et al. [40] showed that a retrieval model based on translation probabilities, learned from a question and answer archive, can recognize semantically similar questions. Duan et al. [41] proposed a dedicated language modeling approach for question search, using question *topic* (user's interest) and question *focus* (certain aspect of the topic).

Lately, these efforts were supported by a task on Question-Question similarity introduced in the community QA challenge at SemEval (task 3B) [4]. Given a new question, the task focused on reranking all similar questions retrieved by a search engine, assuming that the answers to the similar questions will be correct answers for the new question. Different machine learning and deep learning approaches were tested in the scope of SemEval 2016 [4] and 2017 [5] task 3B. The best performing system in 2017 achieved 47.22% Mean Average Precision (MAP) using supervised logistic regression which combined different unsupervised similarity measures such as Cosine and Soft-Cosine [42]. The second-best system achieved

[3] http://netwellness.org/

46.93% MAP with a learning-to-rank method using logistic regression and a rich set of features including lexical and semantic features as well as embeddings generated by different neural networks (siamese, Bi-LSTM, GRU and CNNs) [43]. In the scope of this challenge, a dataset was collected from Qatar Living forum for training. We refer to this dataset as *SemEval-cQA*.

In another effort, an answer-based definition of RQE was proposed and tested [2]. The authors introduced a dataset of clinical questions and used a feature-based method that provided an accuracy of 75% on consumer health questions. We will call this dataset *Clinical-QE*[4]. Dos Santos et al. [6] proposed a new approach to retrieve semantically equivalent questions combining a bag-of-words representation with a distributed vector representation created by a CNN and user data collected from two Stack Exchange communities. Lei et al. [8] proposed a recurrent and convolutional model (gated convolution) to map questions to their semantic representations. The models were pre-trained within an encoder-decoder framework. These works showed the potential relevance of neural networks and traditional machine learning methods in the detection of question similarity and entailment. In this paper we conduct experiments comparing both kinds of approaches for the RQE task, using open-domain and medical datasets for training.

In the next section, we present RQE methods and compare their performance using open-domain and clinical datasets. "Building a Medical QA Collection from Trusted Resources" section describes the new collection of medical question-answer pairs. In "The Proposed Entailment-based QA System" section, we describe our RQE-based approach to QA. "Evaluating RQE for Medical Question Answering" section presents our evaluation of the retrieved answers and the results obtained on TREC 2017 LiveQA medical questions.

RQE Approaches and Experiments

The choice of two methods for our empirical study is motivated by the best performance achieved by logistic regression in question-question similarity at SemEval 2017 (best system [42] and second-best system [43]), and the high performance achieved by neural networks on larger datasets such as SNLI [14, 44–46]. We first define the RQE task, then present the two approaches, and evaluate their performance on five different datasets.

Definition

In the context of QA, the goal of RQE is to retrieve answers to a new question by retrieving entailed questions with associated answers. Therefore, we define question entailment as:

- a question A **entails** a question B if every answer to B is also a **complete** or **partial** answer to A.

We present below two examples of consumer health questions *Ai* and entailed questions *Bi*:
Example 1 (each answer to the entailed question B1 is a *complete* answer to A1):

- A1: *What is the latest news on tennitis, or ringing in the ear, I am 75 years old and have had ringing in the ear since my mid 5os. Thank you.*
- B1: *What is the latest research on Tinnitus?*

Example 2 (each answer to the entailed question B2 is a *partial* answer to A2):

- A2: *My mother has been diagnosed with Alzheimer's, my father is not of the greatest health either and is the main caregiver for my mother. My question is where do we start with attempting to help our parents w/ the care giving and what sort of financial options are there out there for people on fixed incomes.*
- B2: *What resources are available for Alzheimer's caregivers?*

The inclusion of partial answers in the definition of question entailment also allows efficient relaxation of the contextual constraints of the original question A to retrieve relevant answers from entailed, but less restricted, questions.

Deep Learning Model

To recognize entailment between two questions PQ (premise) and HQ (hypothesis), we adapted the neural network proposed by Bowman et al. [14]. The deep learning (DL) model, presented in Fig. 1, consists of three 600d ReLU layers, a bottom layer taking the concatenated sentence representations as input, and a top layer feeding a softmax classifier. The sentence embedding model sums the Recurrent neural network (RNN) embeddings of its words. The word embeddings are first initialized with pretrained GloVe vectors. This adaptation provided the best performance in previous experiments with RQE data.

GloVe[5] is an unsupervised learning algorithm to generate vector representations for words [47]. Training is performed on aggregated word co-occurrence statistics from a large corpus, and the resulting representations show interesting linear substructures of the word vector space. We use the pretrained common crawl version with 840B tokens and 300d vectors, which are not updated during training.

[4] https://github.com/abachaa/RQE_Data_AMIA2016

[5] https://nlp.stanford.edu/projects/glove

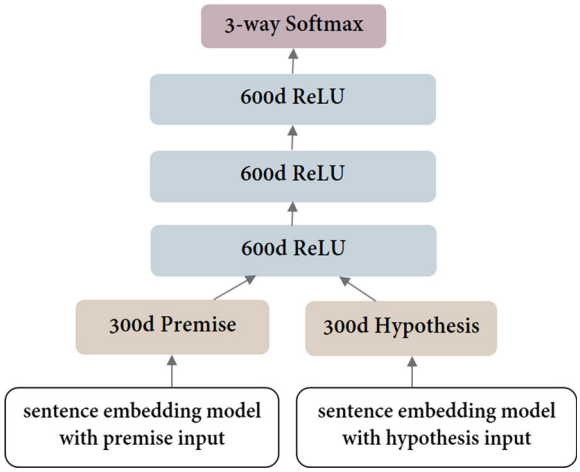

Fig. 1 Neural Network Architecture

Logistic Regression

In this feature-based approach, we use logistic regression to classify question pairs into entailment or no-entailment. Logistic regression achieved good results on this specific task and outperformed other statistical learning algorithms such as SVM and Naive Bayes. In a pre-processing step, we remove stop words and perform word stemming using the Porter algorithm [48] for all (PQ, HQ) pairs.

We use a list of nine features, selected after several experiments on RTE datasets [13]. We compute five similarity measures between the pre-processed questions and use their values as features. We use Word Overlap, the Dice coefficient based on the number of common bigrams, Cosine, Levenshtein, and the Jaccard similarities. Our feature list also includes the maximum and average values obtained with these measures and the question length ratio (length(PQ)/length(HQ)). We compute a morphosyntactic feature indicating the number of common nouns and verbs between PQ and HQ. TreeTagger [49] was used for POS tagging.

For RQE, we add an additional feature specific to the question type. We use a dictionary lookup to map triggers to the question type (e.g. Treatment, Prognosis, Inheritance). Triggers are identified for each question type based on a manual annotation of a set of medical questions (cf. "Question Types" section). This feature has three possible values: 2 (Perfect match between PQ type(s) and HQ type(s)), 1 (Overlap between PQ type(s) and HQ type(s)), and 0 (No common types).

Datasets Used for the RQE Study

Training Datasets

We evaluate the RQE methods (i.e. deep learning model and logistic regression) using two datasets of sentence pairs (SNLI and multiNLI), and three datasets of question pairs (Quora, Clinical-QE, and SemEval-cQA).

The Stanford Natural Language Inference corpus (SNLI) [14] contains 569,037 sentence pairs written by humans based on image captioning. The training set of the MultiNLI corpus [15] consists of 393,000 pairs of sentences from five genres of written and spoken English (e.g. Travel, Government). Two other "matched" and "mismatched" sets are also available for development (20,000 pairs). Both SNLI and multiNLI consider three types of relationships between sentences: entailment, neutral and contradiction. We converted the contradiction and neutral labels to the same non-entailment class.

The Quora dataset of similar questions was recently published with 404,279 question pairs. We randomly selected three distinct subsets (80%/10%/10%) for training (323,423 pairs), development (40,428 pairs) and test (40,428 pairs).

The clinical-QE dataset [2] contains 8588 question pairs and was constructed using 4655 clinical questions asked by family doctors [50]. We randomly selected three distinct subsets (80%/10%/10%) for training (6870 pairs), development (859 pairs), and test (859 pairs).

The question similarity dataset of SemEval 2016 Task 3B (SemEval-cQA) [4] contains 3869 question pairs and aims to rerank a list of related questions according to their similarity to the original question. The same dataset was used for SemEval 2017 Task 3 [5].

RQE Test Dataset

To construct our test dataset, we used a publicly shared set of Consumer Health Questions (CHQs) received by the U.S. National Library of Medicine (NLM), and annotated with named entities, question types, and focus [51, 52]. The CHQ dataset consists of 1721 consumer information requests manually annotated with subquestions, each identified by a question type and a focus.

First, we selected automatically harvested FAQs, from U.S. National Institutes of Health (NIH) websites, that share both the same focus and the same question type with the CHQs. As FAQs are most often very short, we first assume that the CHQ entails the FAQ. Two sets of pairs were constructed: (i) positive pairs of CHQs and FAQs sharing at least one common question type and the question focus, and (ii) negative pairs corresponding to a focus mismatch or type mismatch. For each category of negative examples, we randomly selected the same number of pairs for a balanced dataset. Then, we manually validated the constructed pairs and corrected the positive and negative labels when needed. The final RQE dataset contains 850 CHQ-FAQ pairs with 405 positive and 445 negative pairs. Table 1 presents examples from the five training datasets (SNLI, MultiNLI, SemEval-cQA, Clinical-QE and Quora) and the new test dataset of medical CHQ-FAQ pairs.

Table 1 Description of training and test datasets

Datasets	Type/Domain	# pairs	Positive Examples (Entailment/Similarity)
SNLI (2015)	Inference pairs of open-domain sentences.	550,152 (train)	PS: A child in a light and dark green ensemble sits in a chair in front of a typewriter looking off-camera. HS: A child sitting in front of a desk.
MultiNLI (2017)	Inference pairs of open-domain sentences.	392,702 (train)	PS: On the island of the Giudecca, you'll find another of the great Palladio-designed churches (one of two in Venice), the Redentore. HS: There are two church in Venice that were designed by Palladio.
SemEval-cQA (2016)	Similar questions from the Qatar Living forum.	3169 (train)	PQ: Books. Where can i donate books? HQ: english books. Where to buy english books? Is there a public library in doha? thanks
Clinical-QE (2016)	Entailment pairs of questions asked by doctors.	8588	PQ: Patient is reluctant to take medications so I have been treating with smaller doses than I would with some other patients. How do I control her hypertension and still get her cooperation? HQ: Patient reluctant to take medication. How to control hypertension and still get her cooperation?
Quora (2017)	Open-domain question similarity pairs.	404,279	PQ: I've been working out in the gym for the last three months but I'm not successful in gaining weight. Should I go for a mass gainer? Is it safe? HQ: I have been working out from few months but I am unable to gain mass/weight.Which mass gainer should I take?
New Test Data (CHQs)	Entailment pairs of consumer health questions.	850	PQ: IHSS heart condition and WPW heart condition. Is there any way you could send me information on both these heart conditions? My son has to get tested for them eventually and I would just like information to understand the conditions of both of them more. HQ: What is Wolff-Parkinson-White syndrome ?

Results of RQE Approaches

In the first experiment, we evaluated the DL models and the logistic regression on SNLI, multi-NLI, Quora, and Clinical-QE. For the datasets that did not have a development and test sets, we randomly selected two sets, each amounting to 10% of the data, for test and development, and used the remaining 80% for training. For MultiNLI, we used the dev1-matched set for validation and the dev2-mismatched set for testing. For all DL experiments, the presented results correspond to the best run out of five.

Table 2 presents the results of the first experiment. The DL model with GloVe word embeddings achieved better results on three datasets, with 82.80% accuracy on SNLI, 78.52% accuracy on MultiNLI, and 83.62% accuracy on Quora.

Logistic regression achieved the best accuracy of 98.60% on Clinical-RQE. We also performed a 10-fold cross-validation on the full Clinical-QE data of 8588 question pairs, which gave 98.61% accuracy.

In the second experiment, we used these datasets for training only and compared their performance on our test set of 850 consumer health questions. Table 3 presents the results of this experiment. Logistic regression trained on the clinical-RQE data outperformed DL models trained on all datasets, with 73.18% accuracy.

Table 2 Accuracy (%) of RQE methods using the respective training and test sets of four datasets: SNLI, MultiNLI, Quora, and Clinical-QE

Methods	Textual Datasets		Question Datasets	
	SNLI	MultiNLI	Quora	Clinical-QE
Neural Network (NN)	79.50	73.71	81.34	71.45
NN + GloVe embeddings	**82.80**	**78.52**	**83.62**	93.12
Logistic Regression + Features	75.91	67.88	67.79	**98.60**

The best score are in bold

Table 3 Accuracy (%) of RQE methods when trained using the training sets of SNLI, MultiNLI, Quora and Clinical-QE, and tested on our test set of 850 consumer health questions

Methods	Training Datasets			
	SNLI	MultiNLI	Quora	Clinical-QE
Neural Network (NN)	48.94	54.59	52.35	48.71
NN + GloVe embeddings	49.41	54.82	52.82	57.18
Logistic Regression + Features	67.05	64.94	52.11	**73.18**

The best score are in bold

To validate further the performance of the LR method, we evaluated it on question similarity detection. A typical approach to this task is to use an IR method to find similar question candidates, then a more sophisticated method to select and rerank the similar questions. We followed a similar approach for this evaluation by combining the LR method with the IR baseline provided in the context of SemEval-cQA. The hybrid method combines the score provided by the logistic regression method and the reciprocal rank from the IR baseline using a weight-based combination:

$Hybridscore = LR\_score + w \times \frac{1}{IR\_rank}$

The weight w was empirically set through several tests on the cQA-2016 development set ($w = 8.9$).

Table 4 presents the results on the cQA-2016 and cQA-2017 test datasets. The hybrid method (LR+IR) provided the best results on both datasets. On the 2016 test data, the LR+IR method outperformed the best system in all measures, with 80.57% accuracy and 77.47% MAP (official system ranking measure in SemEval-cQA). On the cQA-2017 test data, the LR+IR method obtained 44.66% MAP and outperformed the cQA-2017 best system in accuracy with 67.27%.

Discussion of RQE Results

When trained and tested on the same corpus, the DL model with GloVe embeddings provided the best results on three datasets (SNLI, MultiNLI and Quora). Logistic regression gave the best accuracy on the Clinical-RQE dataset with 98.60%. When tested on our test set (850 medical CHQs-FAQs pairs), logistic regression trained on Clinical-QE delivered the best performance with 73.18% accuracy. We will investigate other DL solutions in the future using more sophisticated sentence embeddings generated by language models such as BERT or GPT [53–55]. Word embeddings trained specifically for the medical domain can also play an important role in improving the performance [56], as well as different forms of embeddings aggregation that should be investigated when building aggregated sentence embeddings [57, 58].

The SNLI and multi-NLI models did not perform well when tested on the RQE collection of consumer health questions. We performed additional evaluations using the RTE-1, RTE-2 and RTE-3 open-domain datasets provided by the PASCAL challenge and the results were similar. We have also tested the SemEval-cQA-2016 model and had a similar drop in performance on RQE data. This could be explained by the different types of data leading to wrong internal conceptualizations of medical terms and questions in the deep neural layers. This performance drop could also be caused by the complexity of the consumer-health test questions that are often composed of several subquestions, contain contextual information, and may contain misspellings and ungrammatical sentences, which makes them more difficult to process [59]. Another aspect is the semantics of the task as discussed in "Task Definition" section. The definition of textual entailment in open-domain may not quite apply to question entailment in the scope of question answering due to the strict semantics. Also, the general textual entailment definitions refer only to the premise and hypothesis, while the definition of RQE for question answering relies on the relationship between the sets of answers of the compared questions.

Building a Medical QA Collection from Trusted Resources

An RQE-based QA system requires a collection of question-answer pairs to map new user questions to the existing questions with an RQE approach, rank the retrieved questions, and present their answers to the user. We created a collection of medical question-answers pairs, called MedQuAD, that we describe below.

Method

To construct trusted medical question-answer pairs, we crawled websites from the National Institutes of Health[6] (cf. "Medical Resources" section). Each web page describes a specific topic (e.g. name of a disease or a drug), and often includes synonyms of the main topic that we extracted during the crawl.

We constructed handcrafted patterns for each website to automatically generate the question-answer pairs based on the document structure and the section titles. We also annotated each question with the associated focus (topic of the web page), its synonyms (if available in the web page), as well as the question type identified with the designed patterns (cf. "Question Types" section).

To provide additional information about the questions, we performed medical entity recognition to automatically annotate the questions with the focus, its UMLS Concept Unique Identifier (CUI) and Semantic Type. We combined two methods to recognize medical entities from the titles of the crawled articles and their associated UMLS CUIs: (i) exact string matching to the UMLS Metathesaurus[7], and (ii) MetaMap Lite[8] [60]. We then used the UMLS Semantic Network to retrieve the associated semantic types and groups.

These annotations were added to the beginning of the document as shown in Fig. 3:

- Focus: Acromegaly
- UMLS CUI: C0001206
- UMLS Semantic Type: T047
- UMLS Semantic Group: Disorders

[6] www.nih.gov
[7] We used the umls-2017AA version.
[8] https://metamap.nlm.nih.gov/MetaMapLite.shtml

Table 4 Results (%) of the hybrid method (Logistic Regression + IR) on community QA datasets (SemEval-cQA-Test 2016 and SemEval-cQA-Test 2017)

Systems	Test Sets	Acc	P	R	F1	MAP	MRR
Hybrid Method (Logistic Regression + IR)	cQA-16-Test	**80.57**	**70.29**	**72.10**	**71.19**	**77.47**	**83.79**
cQA-B-2016 Best System [4]	cQA-16-Test	*76.57*	*63.53*	*69.53*	*66.39*	*76.70*	*83.02*
cQA-B-2016 IR Baseline [4]	cQA-16-Test	-	-	-	-	*74.75*	*83.79*
Hybrid Method (Logistic Regression + IR)	cQA-17-Test	**67.27**	**33.68**	**79.14**	**47.25**	**44.66**	**48.08**
cQA-B-2017 Best System [5]	cQA-17-Test	*52.39*	*27.30*	*94.48*	*42.37*	*47.22*	*50.07*
cQA-B-2017 IR Baseline [5]	cQA-17-Test	-	-	-	-	*41.85*	*46.42*

The best score are in bold

Followed by the synonyms of the focus when available:

- Somatotroph adenoma
- Growth hormone excess
- Pituitary giant (in childhood)

In this paper, we did not use these additional annotations. They are provided to enrich the MedQuAD dataset for other NLP and IR applications.

Question Types

The question types and their triggers were derived after the manual evaluation of 1721 consumer health questions [51, 52]. Our taxonomy includes 16 types about Diseases, 20 types about Drugs and one type (Information) for the other named entities such as Procedures, Medical exams and Treatments. We describe below the considered question types and examples of associated question patterns.

1 *Types of Questions about Diseases (16)*: Information, Research (or Clinical Trial), Causes, Treatment, Prevention, Diagnosis (Exams and Tests), Prognosis, Complications, Symptoms, Inheritance, Susceptibility, Genetic changes, Frequency, Considerations, Contact a medical professional, Support Groups.
Examples:

   - What research (or clinical trial) is being done for DISEASE?
   - What is the outlook for DISEASE?
   - How many people are affected by DISEASE?
   - When to contact a medical professional about DISEASE?
   - Who is at risk for DISEASE?
   - Where to find support for people with DISEASE?

2 *Types of Questions About Drugs (20):* Information, Interaction with medications, Interaction with food, Interaction with herbs and supplements, Important warning, Special instructions, Brand names, How does it work, How effective is it, Indication, Contraindication, Learn more, Side effects, Emergency or overdose, Severe reaction, Forget a dose, Dietary, Why get vaccinated, Storage and disposal, Usage, Dose.
Examples:

   - Are there interactions between DRUG and herbs and supplements?
   - What important warning or information should I know about DRUG?
   - Are there safety concerns or special precautions about DRUG?
   - What is the action of DRUG and how does it work?
   - Who should get DRUG and why is it prescribed?
   - What to do in case of a severe reaction to DRUG?

3 *Question Type for other medical entities (e.g. Procedure, Exam, Treatment)*: Information.

   - What is Coronary Artery Bypass Surgery?
   - What are Liver Function Tests?

Medical Resources

We used 12 trusted websites to construct a collection of question-answer pairs. For each website, we extracted the free text of each article, as well as the synonyms of the article focus (topic). These resources and their brief descriptions are provided below:

1 National Cancer Institute (NCI) [9]: We extracted free text from 116 articles on various cancer types (729 QA pairs). We manually restructured the content of the articles to generate complete answers (e.g. a full answer about the treatment of all stages of a specific type of cancer). Figure 2 presents examples of QA pairs generated from an NCI article.

2 Genetic and Rare Diseases Information Center (GARD)[10]: This resource contains information about

[9] http://www.cancer.gov/types
[10] https://rarediseases.info.nih.gov/diseases

```
- <Document id="0000065" source="ADAM" url="https://www.nlm.nih.gov/medlineplus/ency/article/000321.htm">
    <Focus>Acromegaly</Focus>
  - <FocusAnnotations>
      <Category>Disease</Category>
    - <UMLS>
      - <CUIs>
          <CUI>C0001206</CUI>
        </CUIs>
      - <SemanticTypes>
          <SemanticType>T047</SemanticType>
        </SemanticTypes>
        <SemanticGroup>Disorders</SemanticGroup>
      </UMLS>
    - <Synonyms>
        <Synonym>Somatotroph adenoma</Synonym>
        <Synonym>Growth hormone excess</Synonym>
        <Synonym>Pituitary giant (in childhood)</Synonym>
      </Synonyms>
    </FocusAnnotations>
  - <QAPairs>
    - <QAPair pid="1">
        <Question qid="0000065-1" qtype="information">What is (are) Acromegaly ?</Question>
      - <Answer>
          Acromegaly is a condition in which there is too much growth hormone in the body.
        </Answer>
      </QAPair>
    - <QAPair pid="2">
        <Question qid="0000065-2" qtype="causes">What causes Acromegaly ?</Question>
      - <Answer>
          Acromegaly is a rare condition. It is caused when the pituitary gland makes too much growth hormone. The pituitary gland is a pea-sized endocrine gland located at the base of the brain. It controls, makes, and releases several hormones, including growth hormone. Usually a noncancerous (benign) tumor of the pituitary gland causes the gland to release too much growth hormone. In children, too much growth hormone causes gigantism rather than acromegaly.
        </Answer>
      </QAPair>
    - <QAPair pid="3">
        <Question qid="0000065-3" qtype="symptoms">What are the symptoms of Acromegaly ?</Question>
      - <Answer>
          Symptoms of acromegaly may include any of the following: - Body odor - Carpal tunnel syndrome - Decreased muscle strength (weakness) - Decreased peripheral vision - Easy fatigue - Excessive height (when excess growth hormone production begins in childhood) - Excessive sweating - Headache - Hoarseness - Joint pain, limited joint movement, swelling of the bony areas around a joint - Large bones of the face - Large feet (change in shoe size), large hands (change in ring or glove size) - Large glands in the skin (sebaceous glands) - Large jaw (prognathism) and tongue (macroglossia) - Sleep apnea - Thickening of the skin, skin tags - Widely spaced teeth - Widened fingers or toes, with swelling, redness, and pain Other symptoms that may occur with this disease: - Colon polyps - Excess hair growth in females (hirsutism) - Type 2 diabetes - Weight gain (unintentional)
        </Answer>
      </QAPair>
    - <QAPair pid="4">
        <Question qid="0000065-4" qtype="exams and tests">How to diagnose Acromegaly ?</Question>
      - <Answer>
          The health care provider will perform a physical exam and ask about your symptoms. The following tests may be ordered to confirm diagnosis of acromegaly: - Blood glucose - Growth hormone - High insulin-like growth factor 1 (IGF-1) level - Spine x-ray - MRI of the brain, including thepituitary gland - Echocardiogram - Prolactin
        </Answer>
      </QAPair>
    - <QAPair pid="5">
        <Question qid="0000065-5" qtype="treatment">What are the treatments for Acromegaly ?</Question>
      - <Answer>
          Surgery to remove the pituitary tumor that is causing this condition often corrects the abnormal growth hormone. Sometimes the tumor is too large to remove completely. People who do not respond to surgery may have radiation of the pituitary gland. Medications are used after surgery. Some patients are treated with medicines instead of surgery. After treatment, you will need to see your health care provider regularly to make sure that the pituitary gland is working normally. Yearly evaluations are recommended.
        </Answer>
```


Fig. 2 Examples of QA pairs generated from an article about *Langerhans Cell Histiocytosis* (NCI)

various aspects of genetic/rare diseases. We extracted all disease question/answer pairs from 4278 topics (5394 QA pairs).

3 Genetics Home Reference (GHR)[11]: This NLM resource contains consumer-oriented information about the effects of genetic variation on human health. We extracted 1099 articles about diseases from this resource (5430 QA pairs).

4 MedlinePlus Health Topics[12]: This portion of MedlinePlus contains information on symptoms, causes, treatment and prevention for diseases, health conditions, and wellness issues. We extracted the free text from the summary sections of 981 articles (981 QA pairs).

5 National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) [13]: We extracted text from 174 health information pages on diseases studied by this institute (1192 QA pairs).

6 National Institute of Neurological Disorders and Stroke (NINDS)[14]: We extracted free text from 277 information pages on neurological and stroke-related diseases from this resource (1104 QA pairs).

[11] https://ghr.nlm.nih.gov
[12] https://medlineplus.gov/healthtopics.html
[13] https://www.niddk.nih.gov/health-information
[14] https://www.ninds.nih.gov/Disorders/all-disorders

```
−<Document id="0000023_1" source="CancerGov" url="https://www.cancer.gov/types/langerhans/patient/langerhans-treatment-pdq">
  <Focus>Langerhans Cell Histiocytosis</Focus>
 −<FocusAnnotations>
  −<UMLS>
   −<CUIs>
      <CUI>C0019621</CUI>
    </CUIs>
   −<SemanticTypes>
      <SemanticType>T191</SemanticType>
    </SemanticTypes>
    <SemanticGroup>Disorders</SemanticGroup>
  </UMLS>
 </FocusAnnotations>
−<QAPairs>
 −<QAPair pid="1">
   <Question qid="0000023_1-1" qtype="information">What is (are) Langerhans Cell Histiocytosis ?</Question>
  −<Answer>
     Key Points - Langerhans cell histiocytosis is a type of cancer that can damage tissue or cause lesions to form in one or more places in the body. - Family
     history or having a parent who was exposed to certain chemicals may increase the risk of LCH. - The signs and symptoms of LCH depend on where it is
     in the body. - Skin and nails - Mouth - Bone - Lymph nodes and thymus - Endocrine system - Central nervous system (CNS) - Liver and spleen - Lung -
     Bone marrow - Tests that examine the organs and body systems where LCH may occur are used to detect (find) and diagnose LCH. - Certain factors
     affect prognosis (chance of recovery) and treatment options. Langerhans cell histiocytosis is a type of cancer that can damage tissue or cause lesions to
     form in one or more places in the body. Langerhans cell histiocytosis (LCH) is a rare cancer that begins in LCH cells (a type of dendritic cell which fight
     infection). Sometimes there are mutations (changes) in LCH cells as they form. These include mutations of the BRAF gene. These changes may make the
     LCH cells grow and multiply quickly. This causes LCH cells to build up in certain parts of the body, where they can damage tissue or form lesions. LCH is
     not a disease of the Langerhans cells that normally occur in the skin. LCH may occur at any age, but is most common in young children. Treatment of
     LCH in children is different from treatment of LCH in adults. The treatments for LCH in children and adults are described in separate sections of this
     summary. Check the list of NCI-supported cancer clinical trials that are now accepting patients with childhood Langerhans cell histiocytosis. For more
     specific results, refine the search by using other search features, such as the location of the trial, the type of treatment, or the name of the drug. Talk
     with your child's doctor about clinical trials that may be right for your child. General information about clinical trials is available from the NCI website
   </Answer>
 </QAPair>
 −<QAPair pid="2">
  −<Question qid="0000023_1-2" qtype="susceptibility">
     Who is at risk for Langerhans Cell Histiocytosis? ?
   </Question>
  −<Answer>
     Anything that increases your risk of getting a disease is called a risk factor. Having a risk factor does not mean that you will get cancer; not having risk
     factors doesn't mean that you will not get cancer. Talk with your doctor if you think you may be at risk. Risk factors for LCH include the following: -
     Having a parent who was exposed to certain chemicals such as benzene. - Having a parent who was exposed to metal, granite, or wood dust in the
     workplace. - A family history of cancer, including LCH. - Having infections as a newborn. - Having a personal history or family history of thyroid diseas
     - Smoking, especially in young adults. - Being Hispanic.
   </Answer>
 </QAPair>
 −<QAPair pid="3">
  −<Question qid="0000023_1-3" qtype="symptoms">
     What are the symptoms of Langerhans Cell Histiocytosis ?
   </Question>
  −<Answer>
     These and other signs and symptoms may be caused by LCH or by other conditions. Check with your doctor if you or your child have any of the
     following: Skin and nails LCH in infants may affect the skin only. In some cases, skin-only LCH may get worse over weeks or months and become a form
     called high-risk multisystem LCH. In infants, signs or symptoms of LCH that affects the skin may include: - Flaking of the scalp that may look like cradle
     cap. - Raised, brown or purple skin rash anywhere on the body. In children and adults, signs or symptoms of LCH that affects the skin and nails may
     include: - Flaking of the scalp that may look like dandruff. - Raised, red or brown, crusted rash in the groin area, abdomen, back, or chest, that may be
     itchy. - Bumps or ulcers on the scalp. - Ulcers behind the ears, under the breasts, or in the groin area. - Fingernails that fall off or have discolored
     grooves that run the length of the nail. Mouth Signs or symptoms of LCH that affects the mouth may include: - Swollen gums. - Sores on the roof of the
     mouth, inside the cheeks, or on the tongue or lips. - Teeth that become uneven. - Tooth loss. Bone Signs or symptoms of LCH that affects the bone may
     include: - Swelling or a lump over a bone, such as the skull, ribs, spine, thigh bone, upper arm bone, elbow, eye socket, or bones around the ear. - Pain
     where there is swelling or a lump over a bone. Children with LCH lesions in bones around the ears or eyes have a high risk for diabetes insipidus and
     other central nervous system disease. Lymph nodes and thymus Signs or symptoms of LCH that affects the lymph nodes or thymus may include: -
     Swollen lymph nodes. - Trouble breathing. - Superior vena cava syndrome. This can cause coughing, trouble breathing, and swelling of the face, neck,
```

Fig. 3 Examples of QA pairs generated from an article about *Acromegaly* (A.D.A.M encyclopedia)

7 NIHSeniorHealth [15]: This website contains health and wellness information for older adults. We extracted 71 articles from this resource (769 QA pairs).

8 National Heart, Lung, and Blood Institute (NHLBI) [16]: We extracted text from 135 articles on diseases, tests, procedures, and other relevant topics on disorders of heart, lung, blood, and sleep (559 QA pairs).

9 Centers for Disease Control and Prevention (CDC) [17]: We extracted text from 152 articles on diseases and conditions (270 QA pairs).

10 MedlinePlus A.D.A.M. Medical Encyclopedia[18]: This resource contains 4366 articles about conditions, tests, and procedures. 17,348 QA pairs were extracted from this resource. Figure 3 presents examples of QA pairs generated from A.D.A.M encyclopedia.

11 MedlinePlus Drugs[19]: We extracted free text from 1316 articles about Drugs and generated 12,889 QA pairs.

12 MedlinePlus Herbs and Supplements[20]: We extracted free text from 99 articles and generated 792 QA pairs.

[15] https://nihseniorhealth.gov/
[16] https://www.nhlbi.nih.gov/health/health-topics
[17] https://www.cdc.gov/diseasesconditions/
[18] https://medlineplus.gov/encyclopedia.html
[19] https://medlineplus.gov/druginformation.html
[20] https://medlineplus.gov/druginfo/herb_All.html

The MedQuAD collection contains 47,457 annotated question-answer pairs about Diseases, Drugs and other named entities (e.g. Tests) extracted from these 12 trusted resources.

The Proposed Entailment-based QA System

Our goal is to generate a ranked list of answers for a given Premise Question *PQ* by ranking the recognized Hypothesis Questions *HQs*. Based on the RQE experiments above ("Results of RQE Approaches" section), we selected logistic regression trained on the clinical-RQE dataset to recognize entailed questions and to rank them with their classification scores.

RQE-based QA Approach

Recognizing question entailment between a given user question and all questions in a large collection is not practical for real-time QA systems. Therefore, we first filter the questions of the MedQuAD dataset with an IR method to retrieve candidate questions, then classify them as entailed (or not) by the user/test question. Based on the positive results of the combination method tested on SemEval-cQA data ("Results of RQE Approaches" section), we adopted a combination method to merge the results obtained by the search engine and the RQE scores. The answers from both methods are then combined and ranked using an aggregate score. Figure 4 presents the overall architecture of the proposed QA system. We describe each module in more details next.

Finding Similar Question Candidates

For each premise question *PQ*, we used the Terrier search engine[21] to retrieve *N* relevant question candidates $\{HQ_j, j \in [1, N]\}$ and then applied the RQE method to predict the labels for the pairs (PQ, HQ_j).

We indexed the questions of our QA collection without the associated answers. In order to improve the indexing and the performance of question retrieval, we also indexed the synonyms of the question focus and the triggers of the question type with each question. This choice allowed us to avoid the shortcomings of query expansion, including incorrect or irrelevant synonyms and the increased execution time. The synonyms of the question focus (topic) were extracted automatically from the QA collection. The triggers of each question type were defined manually in the question types taxonomy. Below are two examples of indexed questions from our QA collection, with the automatically added focus synonyms and question type triggers:

1 What are the treatments for Torticollis?

- Focus: *Torticollis*. Question type: *Treatment*.
- Added focus synonyms: "Spasmodic torticollis, Wry neck, Loxia, Cervical dystonia". Added question type triggers: "relieve, manage, cure, remedy, therapy".

2 What is the outlook for Legionnaire disease?

- Focus: *Legionnaire disease*. Question Type: *Prognosis*.
- Added focus synonyms: "Legionella pneumonia, Pontiac fever, Legionellosis". Added question type triggers: "prognosis, life expectancy".

The IR task consists of retrieving hypothesis questions HQ_j relevant to the submitted question *PQ*. Following the good performance obtained by result fusion techniques at TREC, we merged the results of the TF-IDF weighting function and the In-expB2 DFR model [61].

Let $QL^V = HQ_1^V, HQ_2^V, ..., HQ_N^V$ be the set of *N* questions retrieved by the first IR model *V* and $QL^W = HQ_1^W, HQ_2^W, ..., HQ_N^W$ be the set of *N* questions retrieved by the second IR model *W*. We merge both sets by summing the scores of each retrieved question HQ_j in both QL^V and QL^W lists, then we rerank the hypothesis questions HQ_j.

Combining IR and RQE

The IR models and the RQE method bring different perspectives to the search for relevant candidate questions. In particular, question entailment allows understanding the relations between the important terms, whereas the traditional IR methods identify the important terms, but will not detect if the relations are opposite. Moreover, some of the question types that the RQE method learns will not be deemed as important terms by traditional IR and the most relevant questions will not be ranked at the top of the list.

Therefore, in our approach, when a question is submitted to the system, candidate questions are fetched using the IR models, then the RQE method is applied to filter out the non-entailed questions and rerank the remaining candidates.

Specifically, we denote *CL* the list of question candidates $\{HQ_j, 1 \leq j \leq N\}$ returned by the IR system. The premise question *PQ* is then used to construct N question pairs $\{(PQ, HQ_j), 1 \leq j \leq N\}$. The RQE method is then applied to filter out the question pairs that are not entailed and rerank the remaining pairs.

More precisely, let $EL^{PQ} = \{HQ_1, HQ_2, \ldots, HQ_k \ldots\}$ in *CL* be the list of selected candidate questions that have a positive entailment relation with a given premise question *PQ*. We rank EL^{PQ} by computing a hybrid score $Hscore_k$ for each candidate question HQ_k taking into account the score of the IR system $score_k(IR)$ and the score of the RQE system $score_k(RQE)$.

[21] http://terrier.org

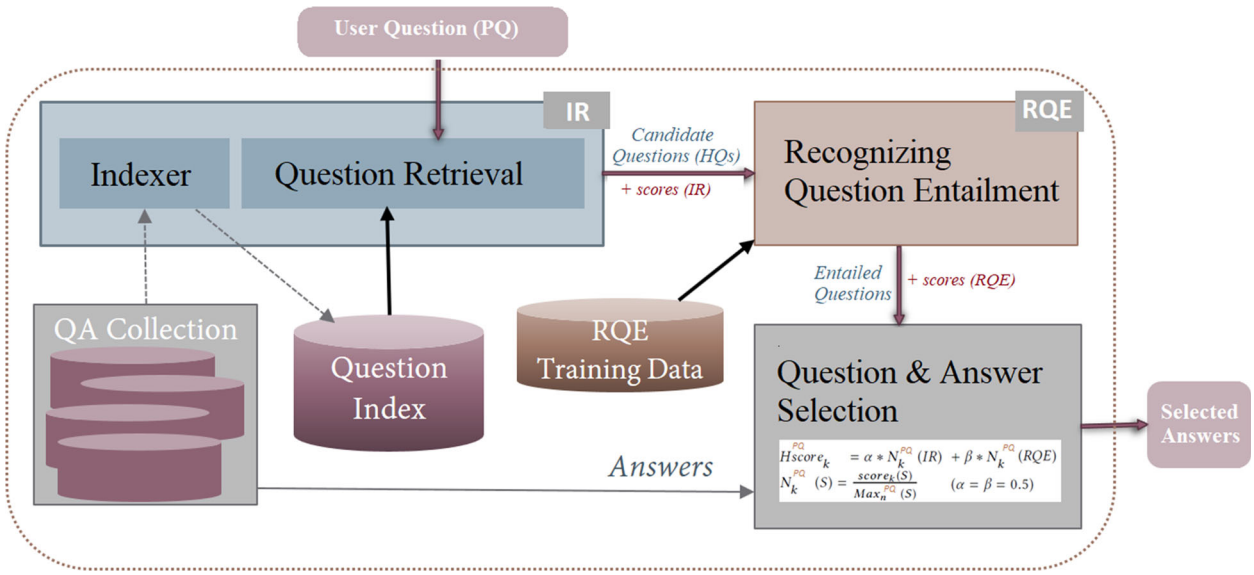

Fig. 4 Overview of the RQE-based Question Answering System

For each system $S \in \{IR, RQE\}$, we normalize the associated score by dividing it by the maximum score among the N candidate questions retrieved by S for PQ:

- $Hscore_k^{PQ} = \alpha * Norm_k^{PQ}(IR) + \beta * Norm_k^{PQ}(RQE)$
- $Norm_k^{PQ}(S) = \frac{score_k(S)}{Max_N^{PQ}(S)} \qquad (\alpha = \beta = 0.5)$

In our experiments, we fixed the value of N to 100. This threshold value was selected as a safe value for this task for the following reasons:

- Our collection of 47,457 question-answer pairs was collected from only 12 NIH institutes and is unlikely to contain more than 100 occurrences of the same focus-type pair.
- Each question was indexed with additional annotations for the question focus, its synonyms, and the question type synonyms.

Evaluating RQE for Medical Question Answering

The objective of this evaluation is to study the effectiveness of RQE for Medical Question Answering, by comparing the answers retrieved by the hybrid entailment-based approach, the IR method, and the other QA systems participating to the medical task at TREC 2017 LiveQA challenge (LiveQA-Med).

Evaluation Method

We developed an interface to perform the manual evaluation of the retrieved answers. Figure 5 presents the evaluation interface showing for each test question, the top ten answers of the evaluated QA method and the reference answer(s) used by LiveQA assessors to help judge the retrieved answers by the participating systems.

We used the test questions[22] of the medical task at TREC-2017 LiveQA [12]. These questions were randomly selected from the consumer health questions that NLM receives daily from all over the world.
The test questions cover different medical entities and have a wide list of question types such as Comparison, Diagnosis, Ingredient, Side effects, and Tapering.

For a relevant comparison, we used the same judgment scores as the LiveQA Track:

- Correct and Complete Answer (4)
- Correct but Incomplete (3)
- Incorrect but Related (2)
- Incorrect (1)

We evaluated the answers returned by the IR-based method and the hybrid QA method (IR+RQE) according to the same reference answers used in LiveQA-Med. The answers were anonymized (the method names were blinded) and presented to 3 assessors: a medical doctor (Assessor A), a medical librarian (B), and a researcher in medical informatics (C). None of the assessors participated in the development of the QA methods. Assessors B and C evaluated 1000 answers retrieved by each of the methods (IR and IR+RQE). Assessor A evaluated 2000 answers from both methods.

Table 5 presents the inter-annotator agreement (IAA) through F1 score computed by considering one of the

[22] https://github.com/abachaa/LiveQA_MedicalTask_TREC2017

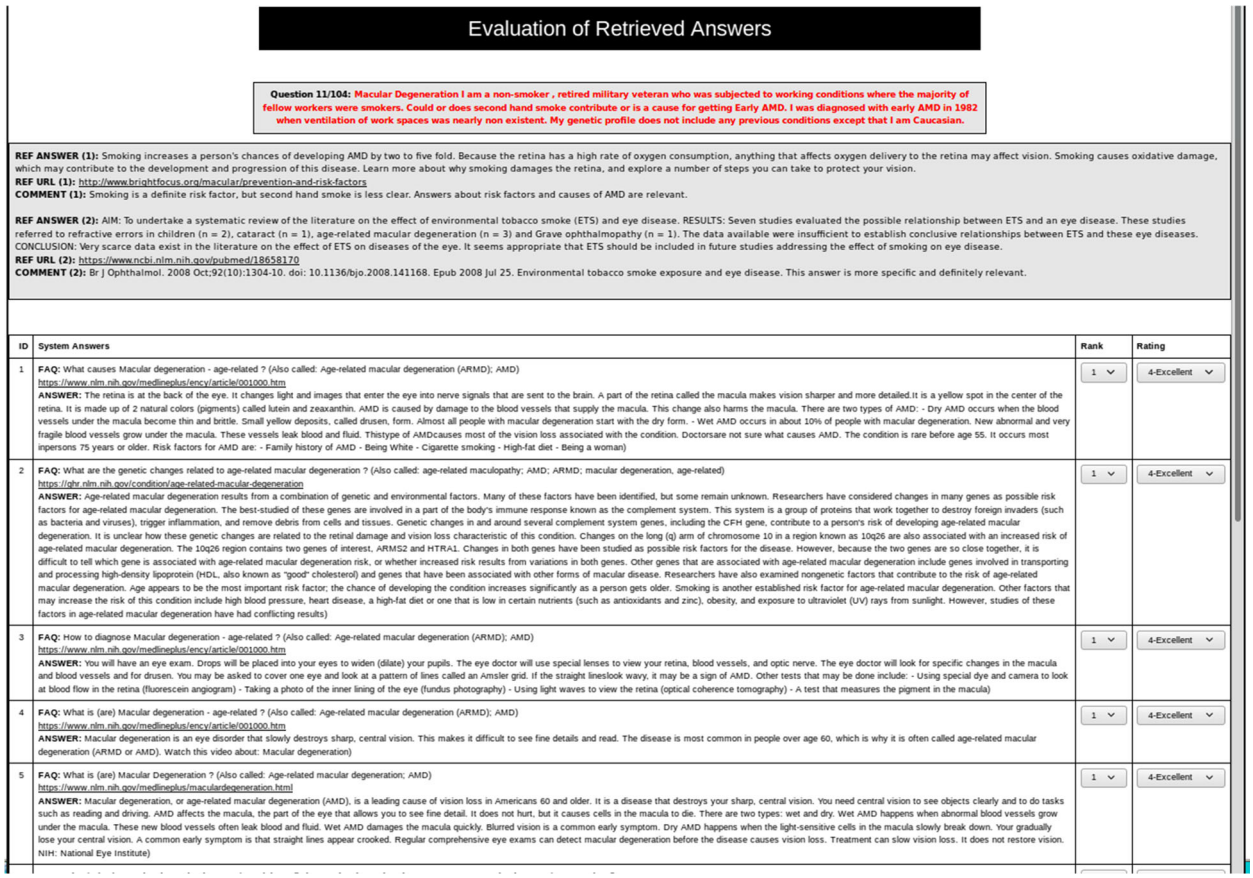

Fig. 5 Evaluation Interface: the reference answers used by LiveQA assessors were provided to help judge the retrieved answers

assessors as reference. In the first evaluation, we computed the True Positives (TP) and False Positives (FP) over all ratings and the Precision and F1 score. As there are no negative labels (only true or false positives for each category), Recall is 100%. We also computed a partial IAA by grouping the "Correct and Complete Answer" and "Correct but Incomplete" ratings (as Correct), and the "Incorrect but Related" and "Incorrect" ratings (as Incorrect). The average agreement on distinguishing the Correct and Incorrect answers is 94.33% F1 score. Therefore, we used the evaluations performed by assessor A for both methods. The official results of the TREC LiveQA track relied on one assessor per question as well.

Table 5 Inter-Annotator Agreement (IAA) over all ratings in the manual evaluation of the retrieved answers

Assessors	IAA		Partial IAA	
	P (%)	F1 (%)	P (%)	F1 (%)
A vs. B	80.80	89.38	90.13	94.81
A vs. C	77.92	87.59	88.42	93.85
Average	79.36	88.48	89.27	94.33

Partial IAA over two ratings "Correct" and "Incorrect"

Evaluation of the first retrieved answer

We computed the measures used by TREC LiveQA challenges [12, 62] to evaluate the first retrieved answer for each test question:

- avgScore(0-3): 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.

Table 6 presents the average scores, success and precision results. The hybrid IR+RQE QA system achieved better results than the IR-based system with 0.827 average score. It also achieved a higher score than the best results

Table 6 LiveQA Measures: Average Score (main score), Success@i+ and Precision@i+ on LiveQA'17 Test Data

Measures	IR-based System	IR+RQE System	LiveQA'17 Best Results	LiveQA'17 Median Results
avgScore(0-3)	0.711	**0.827**	0.637	0.431
succ@2+	0.442	**0.461**	0.392	0.245
succ@3+	0.192	0.25	**0.265**	0.142
succ@4+	0.077	**0.115**	0.098	0.059
prec@2+	0.46	**0.475**	0.404	0.331
prec@3+	0.2	0.257	**0.273**	0.178
prec@4+	0.08	**0.119**	0.101	0.077

Evaluation of the first retrieved answer for each question. N.B. Evaluating the RQE System alone is not relevant as explained previously ("RQE-based QA Approach" section). The best score are in bold

achieved in the medical challenge at LiveQA'17. Evaluating the RQE system alone is not relevant, as applying RQE on the full collection for each user question is not feasible for a real-time system because of the extended execution time.

Evaluation of the top ten answers

In this evaluation, we used Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) which are commonly used in QA to evaluate the top ten answers for each question. We consider answers rated as "Correct and Complete Answer" or "Correct but Incomplete" as correct answers, as the test questions contain multiple subquestions while each answer in our QA collection can cover only one subquestion.

MAP is the mean of the Average Precision (AvgP) scores over all questions.

$$(1)\ MAP = \frac{1}{Q}\sum_{i=1}^{Q} AvgP_i$$

- *Q is the number of questions. $AvgP_i$ is the AvgP of the i^{th} question.*

$$AvgP = \frac{1}{K}\sum_{n=1}^{K} \frac{n}{rank_n}$$

- *K is the number of correct answers. $rank_n$ is the rank of n^{th} correct answer.*

MRR is the average of the reciprocal ranks for each question. The reciprocal rank of a question is the multiplicative inverse of the rank of the first correct answer.

$$(2)\ MRR = \frac{1}{Q}\sum_{i=1}^{Q} \frac{1}{rank_i}$$

- *Q is the number of questions. $rank_i$ is the rank of the first correct answer for the i^{th} question.*

Table 7 presents the MAP@10 and MRR@10 of our QA methods. The IR+RQE system outperforms the IR-based QA system with 0.311 MAP@10 and 0.333 MRR@10.

Discussion of entailment-based QA for the medical domain

In our evaluation, we followed the LiveQA evaluation method with the highest possible rigor. In particular, we consulted with NIST assessors who provided us with the paraphrases of the test questions that they created and used to judge the answers. We used the NIST paraphrases as well as the LiveQA rating approach. Our IAA on the answers rating was also high compared to related tasks, with an 88.5% F1 agreement with the exact four categories and a 94.3% agreement when reducing the categories to two: "Correct" and "Incorrect" answers. Our results show that RQE improves the overall performance and exceeds the best results in the medical LiveQA'17 challenge by a factor of 29.8%.

This performance improvement is particularly interesting as:

(a) Our answer source has only 47K question-answer pairs when LiveQA participating systems relied on much larger collections, including the World Wide Web.
(b) Our system answered one subquestion at most when many LiveQA test questions had several subquestions.

The latter observation, (b), makes the hybrid IR+RQE approach even more promising due to its potential to improve answer completeness.

The former observation, (a), provides another interesting insight: restricting the answer source to only reliable collections can actually improve the QA performance without losing coverage (i.e., our QA approach provided at least one answer to each test question and obtained the best relevance score).

In another observation, the assessors reported that many of the returned answers had a correct question type but a wrong focus, which indicates that including a focus recognition module to filter such wrong answers can improve further the QA performance in terms of precision. Another aspect reported was the repetition of the same (or similar) answer from different websites, which

Table 7 Common Measures: MAP and MRR on LiveQA'17 Test Questions

Measures	IR-based System	IR+RQE System
Fully answered questions	29%	27%
Correctly answered questions	51%	54%
MAP@10	0.282	0.311
MRR@10	0.281	0.333

Evaluation of top 10 answers for each question

could be addressed by improving answer selection with inter-answer comparisons and removal of near duplicates. Also, half of the LiveQA test questions are about Drugs, when only two of our resources are specialized in Drugs, among 12 sub-collections overall. Accordingly, the assessors noticed that the performance of the QA systems was better on questions about diseases than on questions about drugs, which suggests a need for extending our medical QA collection with more information about drugs and associated question types.

We also looked closely at the private websites used by the LiveQA-Med annotators to provide some of the reference answers for the test questions. For instance, the ConsumerLab website was useful to answer a question about the ingredients of a Drug (COENZYME Q10). Similarly, the eHealthMe website was used to answer a test question about interactions between two drugs (Phentermine and Dicyclomine) when no information was found in DailyMed. eHealthMe provides healthcare big data analysis and private research / studies including self-reported adverse drug effects by patients.

However, the question remains on the extent to which big data and other private websites could be used to automatically answer medical questions if information is otherwise unavailable. Unlike medical professionals, patients do not necessarily have the knowledge and tools to validate such information. An alternative approach is putting limitations on medical QA systems in terms of the questions that can be answered (e.g. "What is my diagnosis for such symptoms") and build classifiers to detect such questions and warn the users about the dangers of looking for their answers online.

Generally, medical QA systems should follow strict guidelines regarding the goal and background knowledge and resources of each system, in order to protect the consumers from misleading and/or harmful information. Such guidelines could be based (i) on the source of the information such as health and medical information websites sponsored by the U.S. government, not-for-profit health or medical organizations, and medical university centers, or (ii) on conventions such as the Health on the Net Foundation Code of Conduct (HONcode[23]) that addresses the reliability and credibility of medical and health websites.

Our experiments show that limiting the number of answer sources with such guidelines is not only feasible, but it could also enhance the performance of the QA system from an information retrieval perspective.

Conclusions

In this paper, we carried out an empirical study of recognizing question entailment in the medical domain using several datasets. We developed an RQE-based QA system to answer new medical questions using the existing question-answer pairs. We built and shared a collection of 47K medical question-answer pairs[24]. Our QA approach outperformed the best results on TREC-2017 LiveQA medical test questions. The proposed approach can be applied and adapted to open-domain as well as specific-domain QA. Deep learning models achieved state-of-the-art results on open-domain and clinical datasets, but obtained a lower performance on consumer health questions. We will continue investigating other network architectures including transfer learning, as well as the creation of a large collection of consumer health questions for training to improve the performance of DL models [63]. Future work also includes exploring integration of a Question Focus Recognition module to enhance candidate question retrieval, and, expanding our question-answer collection.

Annex A: Case Study

We consider the case of the question number 36 in the TREC-2017 LiveQA medical test dataset:

36. congenital diaphragmatic hernia. what are the causes of congenital diaphragmatic hernia? Can cousin marriage cause this? What kind of lung disease the baby might experience life long?

This question was answered by 5 participating runs (vs. 8 runs for other questions), and all submitted answers were wrong (scores of 1 or 2). However, our IR-based QA system retrieved one excellent answer (score 4) and our hybrid IR+RQE system provided 3 excellent answers.

A) TREC 2017 LiveQA-Med Participants' Results:

1 36 CMU-OAQA-CarnegieMellonUniversity 2

- 36 2 Yes Opinion 2: CHERUBS is a non-profit organization created to help families affected by Congenital Diaphragmatic Hernia. If you visit the site you will learn all about CDH, read stories of patients (children and adults), see photos and more. Membership and information is free. http://www.cdhsupport.org My name is Dawn and I run CHERUBS. My son was born with CDH and I started CHERUBS 16 years ago. Now, we have over 3600 families in our group so I know there will be someone who truly understands what you're going through. If you'd like to talk or have questions, you can reach me at dawn.williamson@cdhsupport.org

2 36 CMU-LiveMedQA-CarnegieMellonUniversity 1

[23] https://www.hon.ch/HONcode/

[24] https://github.com/abachaa/MedQuAD

- 36 1 Infants with congenital diaphragmatic hernias most commonly present with respiratory distress and cyanosis in the first minutes or hours of life, although a later presentation is possible. The respiratory distress can be severe and may be associated with circulatory insufficiency, requiring aggressive resuscitative measures. ; See Clinical Presentationfor more detail. Clinical Presentation ; As noted in Mortality/Morbidity, population-based studies show that congenital diaphragmatic hernia is diagnosed based on prenatal ultrasonography findings in approximately one half of affected infants. Infants may have a prenatal history ofpolyhydramnios. polyhydramnios Infants most commonly present with respiratory distress and cyanosis in the first minutes or hours of life, although a later presentation is possible. The respiratory distress can be severe and may be associated with circulatory insufficiency, requiring aggressive resuscitative measures. ; Infants with congenital diaphragmati

3 36 ECNU_ICA_2-EastChinaNormalUniversity 1

- 36 1 Congenital Diaphragmatic Hernia (CDH) is a congenital abnormality that is often life-threatening because it can limit the growth of the lungs. The site of herniation is most commonly found on the left side in the hole originally designed for the entry of alimentary canal. Herniation is sometimes found near sternum. Contents * Defects It involves three major defects. * A failure of the diaphragm to completely close during development. * Herniation of the abdominal contents into the chest * Pulmonary Hypoplasia Early diagnosis and presentation This condition can often be diagnosed before birth and fetal intervention can sometimes help, depending on the severity of the condition.[1] Infants born with diaphragmatic hernia experience respiratory failure due to both pulmonary hypertension and pulmonary hypoplasia. The first condition is a restriction of blood flow through the lungs thought to be caused by defects in the lung. Pulmonary hypoplasia or decreased lung volume is directly related

4 36 ECNU-EastChinaNormalUniversity 1

- 36 1 The act of facial motion starts in the brain and travels through the facial nerves to the muscles in the face. These muscles then contract in response to a stimulus. Inside the skull, the facial nerve is a single nerve. Congenital facial palsy is a condition present at birth. Moebius syndrome is a congenital condition. In most cases the exact cause of congenital palsy is uncertain.

5 36 ECNU_ICA-EastChinaNormalUniversity 2

- 36 2 CHERUBS is a non-profit organization created to help families affected by Congenital Diaphragmatic Hernia. If you visit the site you will learn all about CDH, read stories of patients (children and adults), see photos and more. Membership and information is free. http://www.cdhsupport.org My name is Dawn and I run CHERUBS. My son was born with CDH and I started CHERUBS 16 years ago. Now, we have over 3600 families in our group so I know there will be someone who truly understands what you're going through. If you'd like to talk or have questions, you can reach me at dawn.williamson@cdhsupport.org

B) Our IR-based QA System:

- Answer S1-1 (MedlinePlus[25]: What causes Diaphragmatic hernia ? (Also called: Hernia - diaphragmatic; Congenital hernia of the diaphragm (CDH)) ANSWER: A diaphragmatic hernia is a rare defect. It occurs while the baby is developing in the womb. Because the diaphragm is not fully developed, organs, such as the stomach, small intestine, spleen, part of the liver, and the kidney, may take up part of the chest cavity. CDH most often involves only one side of the diaphragm. It is more common on the left side. Often, the lung tissue and blood vessels in the area do not develop normally either. It is not clear if the diaphragmatic hernia causes the underdeveloped lung tissue and blood vessels, or the other way around. 40% of babies with this condition have other problems as well. Having a parent or sibling with the condition increases the risk.

C) Our IR+RQE QA System:

1 Answer S2-1 (from GHR[26]): What are the genetic changes related to congenital diaphragmatic hernia ? (Also called: congenital diaphragmatic defect) ANSWER: Congenital diaphragmatic hernia has many different causes. In 10 to 15 percent of affected individuals, the condition appears as a feature of a disorder that affects many body systems, called a

[25] https://www.nlm.nih.gov/medlineplus/ency/article/001135.htm
[26] https://ghr.nlm.nih.gov/condition/congenital-diaphragmatic-hernia

syndrome. Donnai-Barrow syndrome, Fryns syndrome, and Pallister-Killian mosaic syndrome are among several syndromes in which congenital diaphragmatic hernia may occur. Some of these syndromes are caused by changes in single genes, and others are caused by chromosomal abnormalities that affect several genes. About 25 percent of individuals with congenital diaphragmatic hernia that is not associated with a known syndrome also have abnormalities of one or more major body systems. Affected body systems can include the heart, brain, skeleton, intestines, genitals, kidneys, or eyes. In these individuals, the multiple abnormalities likely result from a common underlying disruption in development that affects more than one area of the body, but the specific mechanism responsible for this disruption is not clear. Approximately 50 to 60 percent of congenital diaphragmatic hernia cases are isolated, which means that affected individuals have no other major malformations. More than 80 percent of individuals with congenital diaphragmatic hernia have no known genetic syndrome or chromosomal abnormality. In these cases, the cause of the condition is unknown. Researchers are studying changes in several genes involved in the development of the diaphragm as possible causes of congenital diaphragmatic hernia. Some of these genes are transcription factors, which provide instructions for making proteins that help control the activity of particular genes (gene expression). Others provide instructions for making proteins involved in cell structure or the movement (migration) of cells in the embryo. Environmental factors that influence development before birth may also increase the risk of congenital diaphragmatic hernia, but these environmental factors have not been identified.

2 Answer S2-2 (from GHR): Is congenital diaphragmatic hernia inherited ? (Also called: congenital diaphragmatic defect): Isolated congenital diaphragmatic hernia is rarely inherited. In almost all cases, there is only one affected individual in a family. When congenital diaphragmatic hernia occurs as a feature of a genetic syndrome or chromosomal abnormality, it may cluster in families according to the inheritance pattern for that condition.

3 Answer S2-3 (from MedlinePlus[27]) What is the outlook for Diaphragmatic hernia ? (Also called: Hernia - diaphragmatic; Congenital hernia of the diaphragm (CDH)): The outcome of surgery depends on how well the baby's lungs have developed. It also depends on whether there are any other congenital problems. Most often the outlook is good for infants who have a sufficient amount of working lung tissue and have no other problems. Medical advances have made it possible for over half of infants with this condition to survive. The babies survived will often have ongoing challenges with breathing, feeding, and growth.

Annex B: List of test questions used in the medical task at TREC LiveQA-Med[28]

1 What are the references with noonan syndrome and polycystic renal disease
2 Gluten information. Re:NDC# 0115-0672-50 Zolmitriptan tabkets 5mg. I have celiac disease & need to know if these contain gluten, Thank you!
3 amphetamine salts 20 mg. are they gluten free
4 vdrl positive. vdrl positive patients please tell me what are he doing . Diagnosis and precaution.
5 how much glucagon. How much glucose is in my GlucaGen HypoKit ? Just curious, I know that there is enough because I have used it. Thank you very much
6 ANESTHESIA EFFECT ON FXTAS PERSONS. Does Anesthesia given during a operation severely hurt, or damage a brain for FXTAS patient? The operation would be for hip replacement! Thank you very much
7 DVT. Can a birth control called Ocella cause DVT? My daughter experiences pains cramping,redness and swelling in her thigh and also really bad huge blood clots during her menstrual cycles after she was prescribed Osella for birth control. Also these syntoms worsened after she gave birth. This has been happening for a year now should she see discuss this with her doctor right away?
8 medication question. I have had a bad UTI for 3 months I have taken cipro 7 times uti returns days after I oomplete I need a new prescription but the doctors here can figure out what to give me as I am allergic to penicillin and allergic to dairy products wich is a filler in many drugs. Please please give me some idea of what I can get my dr; to prescribe
9 can a streptococcus infection cause an invasive disease like wegeners or the symptoms of wegeners?
10 Diabetes and pain control. How can I narrow my search to find information regarding pain(joint) medication suitable to use with a person who has diabetes type 2.
11 Macular Degeneration. I am a non-smoker , retired military veteran who was subjected to working conditions where the majority of fellow workers were smokers. Could or does second hand smoke contribute or is a cause for getting Early AMD. I was

[27] https://www.nlm.nih.gov/medlineplus/ency/article/001135.htm

[28] https://github.com/abachaa/LiveQA_MedicalTask_TREC2017

diagnosed with early AMD in 1982 when ventilation of work spaces was nearly non existent. My genetic profile does not include any previous conditions except that I am Caucasian.

12 molar pregnancy. is conception a requirement of a molar pregnancy. if so, when ?

13 symptoms and diagnosis. My son is being tested now to see if he has hnpp and after reading about the disease, it occurred to me that all my trouble with my hands could have been this and not arthritis. I have had both hands operated on several times, with some success, but continue with swelling in my hands and feet/ankles and soreness and stiffness. Would it be easy to think a patient has arthritis?

14 Yes my wife has been dianosed with giant cell vasculitis Our doctors are not clear about this so im asking for help From you . She has vomited something like coffee grounds and swelling in her feet and legs is really bad.migranes and face swelling to.no blood clots but nothing to go on so please help if u can thank u [NAME] [CONTACT]

15 can't find an answer. I was diagnosed with Fibromyalgia with chronic pain along with some other things and my blood work showed that I was missing a chromosone. How would I find out if I have a genetic for of Fibromyalgia?

16 cant use site. I want to find a doctor who specializes in burning mouth syndrome and that could be in many specialities, I cannot understand how to do this on your website.

17 estradiol 75g patch. Can I stop using the patch only been on it 4.5 months

18 Ear Wax. I sometimes drop Peroxide into the ear and let it bubble for a couple of minutes, then use warm water to flush it out. is there harm?

19 Sevoflurane. 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?

20 ODD. Would like to learn more about condition on ODD

21 Beckwith-Wieddeman Syndrome. I would like to request further knowledge on this specific disorder.

22 CITROBACTOR FREUNDII. Does ciprofaxin work well? Is there a better drug if so what.

23 Hi I have a toddler 22 months and he was long exposure to car seat when he was infant and developed a flat head by then that was resolved, but since then he seems like his back is not well, he only sleep on his tummy, he hates to lay down on his back , he has a bad sitting position when on his car seat and other thing, I was wondering if he may need an evaluation to avoid further damage to his back, please let me know what kinf of doctor should I see, cause his pedi. Dr. Does not has any concerns about it. Thanks.

24 Ear Ache. My son was treated for otitis media on [DATE]. Pains started the previous night. He is taking amoxicillin and antipyrine-benzocaine 5.5%-1.5% ear drops. This morning he woke up with a bit of blood drainage. Is that normal?

25 reaction to these 2 drugs. I would like to know if there are any drug reaction between Carvedilol 25 mg to Hydrslazine50 mg

26 mixing medications. Hello, I would like to know if taking Dicyclomine 20mg, phentermine can have a adverse effect?

27 Dementia. Is dementia genetically passed down or could anyone get it

28 Is there a "sleep apnea surgery". I've heard that there is , but have never found a doctor that does this. My husband has been on C-pap for two years but has not been able to keep it on for more then 2 h. He is not overweight, has had a stroke at 40 years old and double by-pass at 50 years old. Otherwise he follows doctors orders and has no other problems. Thank you for your time, [NAME]

29 Diahrrea. I take Loperamide for chronic diahrrea. Then I stop it for about 2 days so I can have a bowel movement. But then the stool is really soft and there were a few times I almost didn't make to the bathroom. Is there a way for a happy medium

30 about uveitis. IS THE UVEITIS, AN AUTOIMMUNE DISEASE?

31 Customer Service Request. How much urine does an average human bladder hold - in ounces?

32 Amlodipine. I am taking Amlodipine and it has caused my pause rate to be very high. Is there a weaning process when you stop taking Amlodipine and start atenolol? I am taking 5 mg of amlodipine and will be taking 50 mg of atenolol?

33 Shingles vacine. At what age should you get the Shingles shot. My children are in theur late 30's, early 40's, all three had bad cases of chicken pox as children.

34 very worry and need advise . dear sir i had car accident 2 months a go . other person blood splash one me and i saw a lot of blood on my hand and some on face . not sure about eye . i didn't wash it immediatly and until 15 min later then i washed it . am i risk hiv ? thank you .

35 Swan NDC 0869-0871-43. I found 4 cases of expired (04/2010) Hydrogen Peroxide. How do I safely dispose of this product?

36 congenital diaphragmatic hernia. what are the causes of congenital diaphragmatic hernia? Can cousin

marriage cause this? What kind of lung disease the baby might experience life long?

37 shingles. need to know about the work place and someone having shingles, especially while handling food.

38 ClinicalTrials.gov - Question - general information. My question to you is: what is the reason that there is very little attentions is to Antiphosoholipid Syndrome? To find the causes and possibly some type of cure for us who struggle with this auto-immune blood disorder? I guess that since it is female directed (9-1 female to male) that no one important enough has died from APS? Oh, by the way, I'm a 58 year old man!

39 side efectes to methadone. i jest started taking methadone and have confusion my face itches

40 methylprednisolole. unable to fine info on the above med

41 Simvastatin. Why is it recommended that this medicine be taken in the evening? Any harm in taking it in the morning?

42 Prednisone. My husband has been on Prednisone for almost a year for a Cancer treatment he had. He started at 30mg and stayed on 10mg until a couple weeks ago. The prednisone was causing other side effects. He reduced down to 5mg for a couple days and now has been off the prednisone for a week. How long should we expect this drug to stay in his system. He is really experiencing chills/fever/abdominal pain..are these common when coming off this drug? Is there anything else we should expect?

43 Does electrical high voltage shock cause swallowing problems in the near future ??

44 calcitonin salmon nasal spray. I picked up a bottle of above but noted it had NOT been refrigerated for at least the 3 days since Rx was filled. Box and literature state "refrigerate until opened." Pharmacist insisted it was ok "for 30 days" although I said that meant after opening. Cost is $54.08 plus need to know if it will be as effective as should be. Thank you.

45 Schmorl's Nodes. I am trying to obtain information on subject matter.

46 Topic not covered. What exactly is sleep paralysis?

47 Article on Exercise for Impaired - Overweight - Asthmatics. I just found the site through the article on breathing difficulty. My frustration is, WHAT exercises to do with asthma? Today I walked out of the door of the house to take a walk, a beautiful, cool, blustery, sunny day. Suddenly I couldn't catch my breath, my upper chest felt 'heavy', and I had to go back inside and sit down for a while. I'm no exercise weenie, before a couple of bad accidents (car crashes waiting at stop lights!) I used to play softball, volleyball, basketball, even a bit of rugby, I was a dancer and a weight-exerciser, a bicyclist, rollerblader, tree climber. Even today, BP is typically 110-120 over 70-80, heart rate is fine too. Is this just asthma? WHAT exercises can I do, safely? Sure, we ALL get it, exercise is good for us. Just which ones? LOTS of us with asthma would love help here. Thanks.

48 hi. I can't find this I take Ambien every night for sleep. I want to no how long before I go to bed am I supposed to take it.

49 bundle blockage. could you please tell me what a bundle blockage is. what are the symptoms. what is usually done for this? Thank you

50 I have an infection in gums...dentist prescribed Cephalexin 500mg...Is this ok to take even though I am ALLERGIC TO PENICILLAN?

51 Arrhthmia. can arrhythmia occurs after ablation? What is the success rate of Ablation? During my Holter test it was found that my Heart rate fluctuates from 254 to 21. How do you rate the situation?

52 fildena. Hello I was wondering if fildena is truly like Viagra. I'm trying to find an alternative since my insurance no longer will cover Viagra for what ever reason. Would like to know all relavent information regarding fildena. About all I've found is that it is not fda approved so any information would be helpful thanks

53 Nph. I am interested in a movement class. I have nph and can find no help with exercise or support group. Any ideas from [LOCATION] Med?

54 ClinicalTrials.gov - Question - specific study. Can Low dose naltrexone be used for treatment of severe depression?

55 How do you catch hepatitis?

56 Jock Itch. I have Jock itch, and I have read through your symptoms. I wanted know if small lumps under the skin around the scrotum area is a symptoms as well? Should I be concerned?

57 What are the causes of rib cage pain? And and the remedy

58 The hantavirus can lead to death?

59 Is there always elevated temperature associated with appendicitis?

60 Frequency. My urologist has prescribed Oxybutinin, 5 mg tablets, NOT in the ER version. I understood they were to be taken once a day, but he has prescribed twice. Is this the correct recommended dosage, or should the prescription have been for once a day?

61 I have a question for your website an it seems to have difficulty answering . I want to know if you take Gabamentine an hydrocodene together what would happen? ; if I take them separately it don't work.

62 Drug interactions. is it safe to take diclofenac when taking lisinopril or aleve or extra-strength Tylenol?

63 patau sydrome/ trisomy 13. i was wondering the condition of trisomy progresses over time (gets worse as they become older) also, how to diognose the disorder thank you!
64 quinine in seltzer water. Is it ok to drink quinine in seltzer water to ease leg cramps? If so, what would be the correct "dosage"? It has a nasty taste but it does ease leg cramps. Thank you.
65 Mite Infestation. Please inform me of the recommended treatment and prevention protocol for mite infestation in humans, particularly one that is non-toxic or has minimal side effects.
66 meds taken with wine at dinnertime. Is it safe to take my meds with wine at dinnertime?
67 neo oxy. pkease send me the indication and usage info for this powder. NEO-OXY 100/100 MR - neomycin sulfate and oxytetracycline hydrochloride powder
68 What are the reasons for Hypoglycemia in newborns.. and what steps should a pregnent take to avoid this.
69 diverticulitis. can diverticulosis or diverticulitis be detected by a cat scan if there is no infection at that time?
70 CAUSE OF A COLD. i UNDERSTAND CONTAGION AND TRANSFERRENCE OF COLD 'GERMS' WHY ARE SOME PEOPLE AFFECTED AND OTHERS NOT?
71 Janumet XR 50mg/1000mg- 1 daily. Doctor prescribed for type 2 diabetes w/Metformin 500 mg 2 times daily. Pharmacy refused to fill stating overdose of Metformin. Who is right & what is maximum daily dosage of Metformin? Pharmacy is a non-public pharmacy for a major city employer plan provided for employees only.
72 SSPE. My son is 33years of age and did not have the measles vaccination.Could SSPE occur at this age or in the future?
73 ClinicalTrials.gov - Question - general information. My granddaughter was born with Klippel-Tranaunay Syndrome...There is very little information about this. We are looking for the current research and treatments available. She is 5 months old now and her leg seems to be most affected. We want to get her help as soon as possible to address the symptoms and treat her condition.
74 Iron Overdose. Um...i took 25 iron pills...what do i do...this was last night
75 Inherited Ricketts. Mother has inherited ricketts. Passing A child but not B child. How likely would B child pass it on their child?
76 Medicare Part B coverage. I suffer with acute fibromyalgia (sp?) and the various drugs my doctor has prescribed for me have little if any effect in helping to control the pain. My doctor has since given me a prescription to have massage therapy which she thought medicare would cover. However, when checking with medicare, it turns out that it does not! Can you suggest any other type of treatment?
77 Homozygout MTHFR A1298C Health Issues and long term prognosis? What is your position on Homozygout MTHFR A1298C Health Issues and long term prognosis?
78 Vitamin D intake. Can high doses of Vitamin D (50,000 IUs per week) cause flatulence, among other possible effects? And is such a high dose safe to raise very low levels of Vitamin D in the body?
79 Shingles. I am looking for information on how to prevent a shingles outbreak.
80 my father age 65 his always leg pain which use medicine
81 CVID. I have recently been diagnosed with CVID. As a person with thyroid a thyroid tumor greater than 1 cm. and several thyroid cysys I am concerned about cancer. What are the current stats. The tumor is being monitored by my endocrinologist.
82 diabete. whats diabete
83 wellbutrin xl 150. how to taper off
84 Periodic liver tests for patients on Lipitor. I was told at one point that anyone on Lipitor should have blood screening for liver damage every 6 months. Is this currently still the recommendation? NOTE: Although I am in recovery, I also have a history of alcoholism.
85 Hi I have heard in order to get benefit of calcium, it should take with Magnesium, is that right ? I bought Calcium Castco (Kirkland) brand with D please let me know if is good for me because I am osteoporosis . Please help me . thanks
86 Testing for EDS. 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.
87 Consultation. Hello! I have acute chronic cervicitis caused by tampon use. It took a year and a half of treatment (medicines, cauterization), but the symptoms do not stop inflammation and analyzes not determined that bacteria produce inflammation me. I wonder if, despite not being a sexual cervicitis infection, it can cause infertility. And that's what makes a tampon that causes inflammation. Thank you very much!
88 trisomy 7. i am a 32 y/o who has had 4 miscarriages in the past 19 months. Upon my last DNC two weeks ago revealed a genetics study diagnosis of the baby having trisomy 7. could you offer me any information on this? could this have been maternal or paternal? is

this something i would be a carrier of? What are the causes? i have tried but haven't found much information

89 metformin. Does metformin cause high blood pressure?

90 aclidinium. is this a steroid? is there a problem using this if there is a possibility of the need for cataract surgery within the next 12 months?

91 intestines digestion and sbsorption. kindly explain the general effects of smoking or rather the effects of nicotine to digestion and arbsoption

92 sswollen feet ankles legs I have fibromyalgia. When suffering from fibromyalgia will that cause swollen in your body . The swollen started yesterday

93 Can cancer spread through blood contact. 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.

94 Plantar Fasiciitis. Is it true that more likely than not that Plantar fasiciitis could be aggravated by a consistancy of weight bearing activities? Are there other forms of aggravation? if so will you please inform me.

95 CAN LIPNODES AND OR LIVER CANCER BE DETECTED IN A UPPER GI. CAN LIPNODES AND OR LIVER CANCER BE DETECTED IN A UPPER GI

96 abscess teeth. Can an abscess teeth cause a heart attack

97 ischemic syncope stroke diagonses. define?

98 SPECIFY COMPONENTS. COENZYME Q10(100-mg). WHAT ARE THE COMPONENTS OF THIS MEDICINE? IS IT USEABLE FOR MUSLIMS?

99 Autoimmune illness. What doctor specializes in testing for and treatment of autoimmune illness?

100 how does effextor cause ED and what is the mimimum amount that causes ED. I take effexor. Is there a mimimum amount that will not cause ED

101 NSAIDS as a potential cause of ED. How long has this non prescription drug been implicated in erectile dysfunction?

102 i want to know more about aeortic stenosis

103 What can cause white cells ti uprate

104 Glimepiride Storage & Alloweable Excursion Data. Can you please provide Glimepiride storage & allowable temperature excursion data, specifically for pharmacy and warehouse storage

Abbreviations

AvgP: Average precision; CHQs: Consumer health questions; CNNs: Convolutional neural networks; CUI: Concept unique identifier; DL: Deep learning; FAQs: Frequently asked questions; HQ: Hypothesis question; IR: Information retrieval; LSTM: Long short term memory; MAP: Mean average precision; MRR: Mean reciprocal rank; NIH: National institutes of health; NLI: Natural language inference; NLM: National library of medicine; PQ: Premise question; QA: Question answering; RNNs: Recurrent neural networks; RQE: Recognizing question entailment; RTE: Recognizing textual entailment

Acknowledgements

We thank Halil Kilicoglu (NLM/NIH) for his help with the crawling and the manual evaluation, Sonya E. Shooshan (NLM/NIH) for her help with the judgment of the retrieved answers, and Ellen Voorhees (NIST) for her help with the TREC LiveQA evaluation. We also thank the reviewers for their valuable feedback and relevant comments and suggestions.

Authors' contributions

AB designed and implemented the RQE methods, performed the deep learning experiments, created the QA collection, conceived the IR and RQE-based QA methods, implemented the QA systems and the evaluation interface, performed the QA results, and wrote the manuscript.
D.D. participated in the design and creation of the QA collection and the selection of the question-type taxonomy and patterns, evaluated manually the retrieved answers by both QA systems, and edited the manuscript. Both authors read and approved the final manuscript.

Funding

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

Availability of data and materials

- Our newly introduced medical question-answer dataset, MedQuAD, is available at: https://github.com/abachaa/MedQuAD
- The TREC LiveQA-Med training and test datasets and the judged answers (qrels) are available at: https://github.com/abachaa/LiveQA_MedicalTask_TREC2017
- The RQE training set used in the final QA system is available at: https://github.com/abachaa/RQE_Data_AMIA2016
- The RQE-based QA approach is a module of the medical QA system CHiQA [64] available online at: https://chiqa.nlm.nih.gov

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 13 March 2019 Revised: 29 September 2019 Accepted: 1 October 2019
Published online: 22 October 2019

References

1. Russell-Rose T, Chamberlain J. Expert Search Strategies: The Information Retrieval Practices of Healthcare Information Professionals. JMIR Med Inform. 2017;5(4):e33. Available from: http://medinform.jmir.org/2017/4/e33/. Accessed 15 Oct 2019.
2. Ben Abacha A, Demner-Fushman D. Recognizing Question Entailment for Medical Question Answering. In: AMIA 2016, American Medical Informatics Association Annual Symposium, Chicago, IL, USA, November 12-16, 2016; 2016. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333286. Accessed 15 Oct 2019.
3. Luo J, Zhang GQ, Wentz S, Cui L, Xu R. SimQ: Real-Time Retrieval of Similar Consumer Health Questions. J Med Internet Res. 2015;17(2):e43. Available from: https://doi.org/10.2196/jmir.3388.
4. Nakov P, Màrquez L, Moschitti A, Magdy W, Mubarak H, Freihat AA, et al. SemEval-2016 Task 3: Community Question Answering; 2016. p. 525–45. Available from: http://aclweb.org/anthology/S/S16/S16-1083.pdf. Accessed 15 Oct 2019.
5. Nakov P, Hoogeveen D, Màrquez L, Moschitti A, Mubarak H, Baldwin T, et al. SemEval-2017 Task 3: Community Question Answering. In: Proceedings of the 11th International Workshop on Semantic Evaluation SemEval '17. Vancouver, Canada: Association for Computational Linguistics; 2017. https://doi.org/10.18653/v1/s17-2003.
6. dos Santos CN, Barbosa L, Bogdanova D, Zadrozny B. Learning Hybrid Representations to Retrieve Semantically Equivalent Questions. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on

Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers; 2015. p. 694–9. Available from: http://aclweb.org/anthology/P/P15/P15-2114.pdf. Accessed 15 Oct 2019.
7. Romeo S, Martino GDS, Barrón-Cedeño A, Moschitti A, Belinkov Y, Hsu W, et al. Neural Attention for Learning to Rank Questions in Community Question Answering. In: COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, Osaka, Japan, December 11-16, 2016; 2016. p. 1734–45. Available from: http://aclweb.org/anthology/C/C16/C16-1163.pdf. Accessed 15 Oct 2019.
8. Lei T, Joshi H, Barzilay R, Jaakkola TS, Tymoshenko K, Moschitti A, et al. Semi-supervised Question Retrieval with Gated Convolutions. In: NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016; 2016. p. 1279–89. Available from: http://aclweb.org/anthology/N/N16/N16-1153.pdf. Accessed 15 Oct 2019.
9. Harabagiu SM, Hickl A. Methods for Using Textual Entailment in Open-Domain Question Answering. In: ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006; 2006. Available from: http://aclweb.org/anthology/P06-1114. Accessed 15 Oct 2019.
10. Negri M, Kouylekov M. Question Answering over Structured Data: an Entailment-Based Approach to Question Analysis. In: Proceedings of the International Conference RANLP-2009. Association for Computational Linguistics; 2009. p. 305–11. Available from: http://www.aclweb.org/anthology/R09-1056. Accessed 15 Oct 2019.
11. Çelikyilmaz A, Thint M, Huang Z. A Graph-based Semi-Supervised Learning for Question-Answering. In: ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore; 2009. p. 719–27. Available from: http://www.aclweb.org/anthology/P09-1081. Accessed 15 Oct 2019.
12. Ben Abacha A, Agichtein E, Pinter Y, Demner-Fushman D. Overview of the Medical Question Answering Task at TREC 2017 LiveQA. In: Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017; 2017. Available from: https://trec.nist.gov/pubs/trec26/papers/Overview-QA.pdf. Accessed 15 Oct 2019.
13. Dagan I, Roth D, Sammons M, Zanzotto FM. Recognizing Textual Entailment: Models and Applications. Synthesis Lectures on Human Language Technologies: Morgan & Claypool Publishers; 2013. https://doi.org/10.2200/s00509ed1v01y201305hlt023.
14. Bowman SR, Angeli G, Potts C, Manning CD. A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics; 2015. https://doi.org/10.18653/v1/d15-1075.
15. Williams A, Nangia N, Bowman SR. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers); 2018. p. 1112–22. Available from: https://www.aclweb.org/anthology/N18-1101.pdf. Accessed 15 Oct 2019.
16. Nangia N, Williams A, Lazaridou A, Bowman S. The RepEval 2017 Shared Task: Multi-Genre Natural Language Inference with Sentence Representations. Association for Computational Linguistics. 2017. Available from: http://www.aclweb.org/anthology/W17-5301. Accessed 15 Oct 2019.
17. Dai Z, Li L, Xu W. CFO: Conditional Focused Neural Question Answering with Large-scale Knowledge Bases. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers; 2016. Available from: http://aclweb.org/anthology/P/P16/P16-1076.pdf. Accessed 15 Oct 2019.
18. Höffner K, Walter S, Marx E, Usbeck R, Lehmann J, Ngomo AN. Survey on challenges of Question Answering in the Semantic Web. Semantic Web. 2017;8(6):895–920. Available from: http://doi.org/10.3233/SW-160247.
19. Athenikos SJ, Han H. Biomedical Question Answering: A Survey. Comput Methods Prog Biomed. 2010;99(1):1–24. Available from: http://dx.doi.org/10.1016/j.cmpb.2009.10.003.
20. Ben Abacha A, Zweigenbaum P. MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. Inf Process Manag J. 2015;51(5):570–94. Available from: http://doi.org/10.1016/j.ipm.2015.04.006.
21. Yang Y, Yu J, Hu Y, Xu X, 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. https://arxiv.org/pdf/1711.05789.pdf.
22. Lally A, Prager JM, McCord MC, Boguraev B, Patwardhan S, Fan J, et al. Question analysis: How Watson reads a clue. IBM J Res Dev. 2012;56(3):2. Available from: http://doi.org/10.1147/JRD.2012.2184637.
23. Ben Abacha A, Zweigenbaum P. Medical question answering: translating medical questions into sparql queries. In: ACM International Health Informatics Symposium, IHI '12, Miami, FL, USA, January 28-30, 2012; 2012. p. 41–50. Available from: http://doi.acm.org/10.1145/2110363.2110372.
24. Mrabet Y, Kilicoglu H, Roberts K, 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 12-16, 2016; 2016. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333243. Accessed 15 Oct 2019.
25. Momtazi S. Unsupervised Latent Dirichlet Allocation for supervised question classification. Inf Process Manag. 2018;54(3):380–93. Available from: http://www.sciencedirect.com/science/article/pii/S0306457318300153. Accessed 15 Oct 2019.
26. Wang M, Smith NA, Mitamura T. What is the Jeopardy Model? A Quasi-Synchronous Grammar for QA. In: EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic; 2007. p. 22–32. Available from: http://www.aclweb.org/anthology/D07-1003. Accessed 15 Oct 2019.
27. Surdeanu M, Ciaramita M, Zaragoza H. Learning to Rank Answers to Non-Factoid Questions from Web Collections. Comput Linguist. 2011;37(2):351–83. Available from: http://doi.org/10.1162/COLI_a_00051.
28. Tymoshenko K, Moschitti A. Assessing the Impact of Syntactic and Semantic Structures for Answer Passages Reranking. In: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015; 2015. p. 1451–60. Available from: http://doi.acm.org/10.1145/2806416.280649.
29. Severyn A, Moschitti A. Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015; 2015. p. 373–82. Available from: https://doi.org/10.1145/2766462.2767738.
30. Zhang H, Rao J, Lin JJ, Smucker MD. Automatically Extracting High-Quality Negative Examples for Answer Selection in Question Answering. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017; 2017. p. 797–800. Available from: https://doi.org/10.1145/3077136.3080645.
31. Jijkoun V, de Rijke M. Retrieving answers from frequently asked questions pages on the web. In: Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 - November 5, 2005; 2005. p. 76–83. Available from: http://doi.acm.org/10.1145/1099554.1099571.
32. Wang K, Ming Z, Chua T. A syntactic tree matching approach to finding similar questions in community-based qa services. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009; 2009. p. 187–94. Available from: http://doi.acm.org/10.1145/1571941.1571975.
33. Burke RD, Hammond KJ, Kulyukin V, Lytinen SL, Tomuro N, Schoenberg S. Question answering from frequently asked question files: Experiences with the faq finder system. AI Mag. 1997;18(2):57.
34. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993;32:281–91.

35. Wang D, 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, November 15-17, 2017; 2017. Available from: https://trec.nist.gov/pubs/trec26/papers/CMU-OAQA-QA.pdf. Accessed 15 Oct 2019.
36. Datla VV, Arora TR, Liu J, Adduru V, Hasan SA, Lee K, et al. Open domain real-time question answering based on asynchronous multiperspective context-driven retrieval and neural paraphrasing. In: Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017; 2017. Available from: https://trec.nist.gov/pubs/trec26/papers/prna-QA.pdf. Accessed 15 Oct 2019.
37. Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016; 2016. p. 2383–92. Available from: http://aclweb.org/anthology/D/D16/D16-1264.pdf. Accessed 15 Oct 2019.
38. Wiese G, Weissenborn D, Neves ML. Neural Question Answering at BioASQ 5B. In: BioNLP 2017, Vancouver, Canada, August 4, 2017; 2017. p. 76–9. Available from: https://doi.org/10.18653/v1/W17-2309.
39. An W, Chen Q, Tao W, Zhang J, Yu J, Yang Y, et al. 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, November 15-17, 2017; 2017.
40. Jeon J, Croft WB, Lee JH. Finding Similar Questions in Large Question and Answer Archives. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management. CIKM '05. New York: ACM; 2005. p. 84–90. Available from: http://doi.acm.org/10.1145/1099554.1099572.
41. Duan H, Cao Y, Lin C, Yu Y. Searching Questions by Identifying Question Topic and Question Focus. In: ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA; 2008. p. 156–64. Available from: http://www.aclweb.org/anthology/P08-1019. Accessed 15 Oct 2019.
42. Charlet D, Damnati G. SimBow at SemEval-2017 Task 3: Soft-Cosine Semantic Similarity between Questions for Community Question Answering. In: Proceedings of the 11th International Workshop on Semantic Evaluation, SemEvalACL 2017, Vancouver, Canada, August 3-4, 2017; 2017. p. 315–9. Available from: http://doi.org/10.18653/v1/S17-2051.
43. Goyal N. LearningToQuestion at SemEval 2017 Task 3: Ranking Similar Questions by Learning to Rank Using Rich Features. In: Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017; 2017. p. 310–4. Available from: http://doi.org/10.18653/v1/S17-2050.
44. Kim S, Hong JH, Kang I, Kwak N. Semantic Sentence Matching with Densely-connected Recurrent and Co-attentive Information. CoRR. 2018. Available from: http://arxiv.org/pdf/1805.11360.pdf.
45. Chen Q, Zhu X, Ling ZH, Inkpen D, Wei S. Neural Natural Language Inference Models Enhanced with External Knowledge. CoRR. 2018. Available from: http://arxiv.org/pdf/1711.04289.pdf.
46. Ghaeini R, Hasan SA, Datla V, Liu J, Lee K, Qadir A, et al. DR-BiLSTM: Dependent Reading Bidirectional LSTM for Natural Language Inference; 2018. Available from: http://arxiv.org/pdf/1802.05577.pdf.
47. Pennington J, Socher R, Manning CD. GloVe: Global Vectors for Word Representation. In: Empirical Methods in Natural Language Processing (EMNLP); 2014. p. 1532–43. Available from: http://www.aclweb.org/anthology/D14-1162. Accessed 15 Oct 2019.
48. Porter M. An Algorithm for Suffix Stripping. Program. 1980;14(3):130–7.
49. Schmid H. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of the International Conference on New Methods in Language Processing. Manchester; 1994. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf.
50. Ely JW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, Pifer EA, et al. A taxonomy of generic clinical questions: classification study. Br Med J. 2000;321:429–32.
51. Kilicoglu H, Ben Abacha A, Mrabet Y, Roberts K, Rodriguez L, Shooshan SE, et al. Annotating Named Entities in Consumer Health Questions. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28; 2016. Available from: http://www.lrec-conf.org/proceedings/lrec2016/summaries/727.html. Accessed 15 Oct 2019.
52. Kilicoglu H, Ben Abacha A, Mrabet Y, Shooshan SE, Rodriguez L, Masterton K, et al. Semantic annotation of consumer health questions. BMC Bioinformatics. 2018;19(1):34:1–34:28. Available from: https://doi.org/10.1186/s12859-018-2045-1.
53. Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017; 2017. p. 670–80. Available from: https://www.aclweb.org/anthology/D17-1070.pdf. Accessed 15 Oct 2019.
54. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers); 2019. p. 4171–86. Available from: https://aclweb.org/anthology/papers/N/N19/N19-1423/. Accessed 15 Oct 2019.
55. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving Language Understanding by Generative Pre-Training. 2018. Available from: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf. Accessed 15 Oct 2019.
56. Chiu B, Crichton GKO, Korhonen A, Pyysalo S. How to Train good Word Embeddings for Biomedical NLP. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing BioNLP@ACL 2016, Berlin, Germany, August 12, 2016; 2016. p. 166–174. https://doi.org/10.18653/v1/W16-2922.
57. Arora S, Liang Y, Ma T. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings; 2017. Available from: https://openreview.net/forum?id=SyK00v5xx. Accessed 15 Oct 2019.
58. Wieting J, Bansal M, Gimpel K, Livescu K. Towards Universal Paraphrastic Sentence Embeddings. In: 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings; 2016. Available from: http://arxiv.org/abs/1511.08198.
59. Roberts K, Demner-Fushman D. Interactive use of online health resources: a comparison of consumer and professional questions. JAMIA. 2016;23(4):802–11. Available from: http://dx.doi.org/10.1093/jamia/ocw024.
60. Demner-Fushman D, Rogers WJ, Aronson AR. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. JAMIA. 2017;24(4):841–4. Available from: https://doi.org/10.1093/jamia/ocw177.
61. Ounis I, Amati G, Plachouras V, He B, Macdonald C, Johnson D. Terrier Information Retrieval Platform; 2005. p. 517–9. Available from: http://doi.org/10.1007/978-3-540-31865-1_37.
62. Agichtein E, Carmel D, Pelleg D, Pinter Y, Harman D. Overview of the TREC 2015 LiveQA Track. In: Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015; 2015. Available from: http://trec.nist.gov/pubs/trec24/papers/Overview-QA.pdf. Accessed 15 Oct 2019.
63. Ben Abacha A, Shivade C, Demner-Fushman D. Overview of the MEDIQA 2019 Shared Task on Textual Inference, Question Entailment and Question Answering. In: Proceedings of ACL-BioNLP 2019, Florence, Italy, August 1; 2019. https://www.aclweb.org/anthology/W19-5039.pdf.
64. Demner-Fushman D, Mrabet Y, Ben Abacha A. Consumer Health Information and Question Answering: Helping consumers find answers to their health-related information needs. J Am Med Inform Assoc (JAMIA). 2019.

Publisher's Note