Open domain real-time question answering based on asynchronous multiperspective context-driven retrieval and neural paraphrasing

Vivek Datla, Tilak Arora, Joey Liu, Viraj Adduru; Sadid A. Hasan, Kathy Lee, Ashequl Qadir, Yuan Ling, Aaditya Prakash; Oladimeji Farri

Artificial Intelligence Laboratory, Philips Research North America, Cambridge, MA, USA {firstname.lastname, kathy.lee_1}@philips.com

Abstract

The live-QA task involves real user questions, extracted from the stream of most recent questions submitted to the Yahoo Answers (YA) site that has not yet been answered. There are two tracks in the live-QA task, general domain, and medical domain. In general domain, unanswered questions are taken from six categories of real-time vahoo question answering feed, and for the medical domain, they are taken from the consumer health questions asked in NLM forums. The answers given by the system for both general and medical tasks are evaluated by human experts looking into accuracy, readability, and preciseness. The features of our opendomain question answering include question decomposition, question focus identification, context identification, answer retrieval and summarization. The current system builds an asynchronous system which has a multi-perspective view of the question being asked by decomposing the question into multiple smaller questions and identifying answers to subquestions and summarizing the answers. Our system performed close to the median in the live-OA task and ranked second in the medical sub-task.

1 Introduction

Open domain question answering is a challenging task and an open problem in the area of AI and NLP. Our approach to addressing this gap in the

community involves building an automated system that answers an open-domain question using multiple sources after understanding the question and the context in which the question is being asked helps to generate better answers. In the past, researchers have used domain-specific expert systems to answer factoid questions. The performance of the systems for such domain-specific factoid questions is close to that of human experts. The open-domain real-life questions amplify the challenge many folds as natural language is ambiguous, and constructing the answer requires an elaborate understanding of the question being asked, expert domain knowledge, as well as language generation models (Datla et al., 2016). The questions asked are mostly soliciting feeling or opinions regarding a certain life event or task performed.

The open domain real-time question answering task increases the complexity even further as one has answered in less than 60 seconds. Additionally, the answers need to be concise as they are restricted to a limit of 1000 characters. The medical sub-task introduced this year focuses on consumer health question answering received by the U.S. National Library of Medicine (NLM). The questions are less noisy compared to the general domain questions.

2 General Domain Question Answering Task Description

The Live-QA track first started in TREC 2015. This year a new sub-task wherein the questions from the medical domain are asked. The competition runs for 24 hrs, for the general track questions being posted on Yahoo Answers site (after some preliminary cleaning) by the real users are posted on to the participating team's servers registered for the competition. Both the general and medi-

^{*}The author is affiliated with Rochester Institute of Technology(vra2128@rit.edu)

[†]The author is also affiliated with Brandeis University (aprakash@brandeis.edu).

Table 1: Topic categories and no. of sub categories

Topic	#sub topics
Arts & Humanities	10
Beauty & Style	5
Health	10
Home & Garden	6
Pets	8
Sports	30
Travel	27

cal tasks run parallel and are identified by header information provided by the competition organizers.

¹ The questions in the general domain are selected from 7 distinct topics shown in Table 1. As the table indicates, topics for the general domain questions have several sub-categories. The category of a question may overlap with other categories. For example "Why does Labor back this kind of behavior?" is identified by the user as in the topic category "Travel" and to the sub-topic category "Australia". However, from the question, we can understand that it belongs to the category "Politics and Government". Sometimes the questions belong to multiple categories, and the user based on his interest/convenience picks only one topic (Datla et al., 2016).

The questions being asked on Yahoo Answers website are mostly subjective and describe a human experience which is often personal and relevant to the topic. The ability for a machine to replicate human understanding of the topic(s) and biases in a subjective question is challenging. Also, the fact that these questions can represent multiple events and potential causal relationships further complicate the Live-QA task. For example, the question "My fiance hates my dog. He ignores him and always complains. He started calling him names. Should I be worried?" posted in the Pets topic shows three parties (me, my fiance, and my dog) with a mix of interpersonal relationships represented as emotions and actions "hates my dog", "ignores him", "always complains", "started calling him names", and "should I be worried"). The answers given by people for this question include suggestions on personal relationships, pet behaviors, and further questions like "Do you want to stay with the person who is cruel to animals?" (Datla et al., 2016).

As the answers given to the subjective questions indicate that there is no one correct answer and answers provided by users indicate the different focus picked while answering the question. To answer such questions, one needs not only to know about the sentiment and the focus of the question but also needs to know the interactions between various facets of the problem. Given these opendomain questions, the big challenge that we need to address is how to access huge domain knowledge to answer such questions (Datla et al., 2016).

3 Medical Domain sub-task

In addition to the main question answering task, a subtask for medical question answering is introduced this year. The questions for this medical subtask focuses on consumer health question answering. Consumer health questions received by the U.S. National Library of Medicine (NLM) are curated for the competition. The questions are complex and contain more than one sub-questions. Sub-questions are defined according to the focus and the question type. There are 23 question categories defined. They are diagnosed, information, complication, symptom, inheritance, prognosis, treatment, usage, cause, ingredient, contraindication, storage and disposal, genetic changes, association, susceptibility, interaction, tapering, indication, diagnosis, side effects, organization, dosage, and prevention. The question type is defined based on the information being asked by the user. We have organized the question types into a smaller set of categories such as information, medication, diagnosis, treatment, and complication.

For the medical sub-task, we used a slightly different strategy in answering the questions compared to the main task. We used Wikipedia as our knowledge base and used the structure inside Wikipedia to identify the page that most likely contains the answer based on the question asked.

The methodology followed is similar to the process followed in (Datla et al., 2017). We used the structure of Wikipedia to identify the most appropriate section that answers the question based on the classification into 23 categories. For classification of the question into 23 categories, we used the feature extraction algorithm in (Sarker and Gonzalez, 2015).

After identifying the type of the question, we use our in-house medical concept identification to extract the medical terms such as disease names,

¹Yahoo Answers - https://answers.yahoo.com/

Table 2: Question Fields

Category	Topic of the question
Sub-category	sub-topic of the question
Question body	description of the question
Question title	actual question posted

medications, clinical symptoms, etc. mentioned in the context and the question. We use the medical terms and demographic information extracted using our tool as an input into the Knowledge Graph-based clinical diagnosis inference engine described in detail in (Datla et al., 2017). Based on the question type we identify the most likely section in the Wikipedia that holds the answer. If the question has multiple sub-question types, then we answer each sub-question at a time in the same order as given by the user.

4 TREC 2017 Competition

During the TREC competition, a question is pushed every minute by the track organizers onto the server registered with the competition. The question posted on our server is a JSON object with fields shown in Table 2. The category, subcategory, and the question-title fields always have entries, while the other fields can sometimes be empty.

The questions from the topics shown in Table 1 sometimes relate to current events, and hence it becomes particularly challenging to address the questions on personal experiences related to current events, pandemics, or ongoing family issues like marriage, divorce, etc.

The time for answering the question is same for both general domain and medical sub-task. If, after 60 seconds there is no response provided by the server to the competition, then the response is assessed as negative and would be penalized. The systems are ranked on two metrics 1) *success*: a ratio of the aggregated scores of the answers to the total questions asked in the competition and 2) *precision*: a ratio of the aggregated scores of the answers to the total number of the questions answered by the system.

5 System for General Domain Question Answering

The components of our system used for general domain questions is shown in Figure 1. The main

features of our system can be summarized as follows:

- NLP engine to parse question and context and normalize the data.
- Evidence curation engine, which identifies the related sources, metadata based on the user's interest. In the current work we have used Google² and Yahoo³ as our data sources.
- Neural paraphrase identification (Hasan et al., 2016) to pick the most similar questions to the target question.
- Keyword extraction engine (Rose et al., 2010) that normalize the available information ranks the evidence and allows for querying to match the given context and evidence semantically.
- Asynchronous retrieval of candidates for answers for decomposed question and context pairs.
- · Answer generation engine
 - Answer focus text identification.
 - Answer focus word identification based on bi-directional attention model (Seo et al., 2016)
 - Answer phrase generation, by having the components of introduction, answer, and conclusion.
 - Answer alignment concerning the context and decomposed questions

Answer summarization

We had three different combinations of sources for three runs submitted for the competition. For the general domain task, we had Google as our source for Run-1, Yahoo as a source for Run-2 and Run-3 had no preference, used both Google and Yahoo and picked the best answer.

For medical sub-task, we followed a slightly different strategy for answering the questions compared to the general sub-task. As a knowledge source, we have used Wikipedia. For every question, we decomposed it into several sub-questions. The most appropriate wiki page is selected based

²http://www.google.com

³https://answers.yahoo.com/

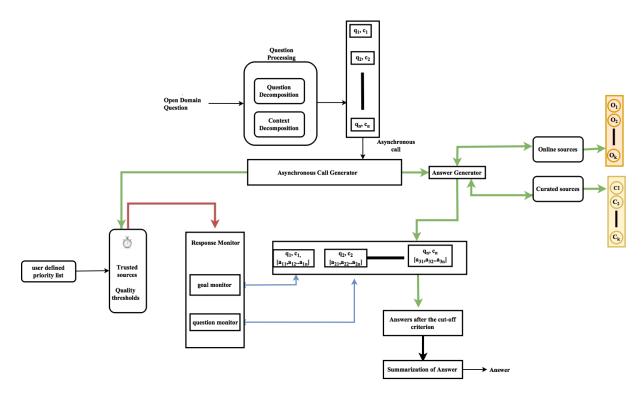


Figure 1: System Architecture for general domain question answering

on the graph structure of Wikipedia. For example, if the focus word is a drug name, then all the pages that have the drug name are retrieved, and the page that has the best fit concerning the description is taken as a candidate source of context from which the answer is extracted. We used an internal medical concept identification engine to capture the medical concepts. We also built a simple bag-of-words based logistic regression classifier to identify the type of the question and the focus of the question based on the description given. For the feature extraction, we have used (Sarker and Gonzalez, 2015).

After identifying the most appropriate page, we do a keyword overlap based windowing to identify the most relevant piece of text that most likely has the answer. This relevant text is used as input to bi-directional attention model (Seo et al., 2016) trained on SQUAD dataset (Rajpurkar et al., 2016). Given a question and context (a body of text that contains the answer), the model picks the phrase that answers the question. We heuristically create the answer by padding the current sentence which has a word with the sentences above and below it. We also take the first sentence of the introduction of that disease as an introductory sentence and prefix the above-generated answer. We had only one system for the medical sub-task. If there

was no answer obtained from using Wikipedia as our source, we defaulted to the sources which we have used in general domain.

6 Results and Discussion

Results from our system were evaluated based on the scoring system shown below:

- avgScore(0-3): The average score over all questions. This is the main score used to rank the participating system runs.
- succ@i+: the number of questions with score i or above ($i \in \{2..4\}$) divided by the total number of questions. For example, succ@2+ measures the percent of questions with at least fair grade answered by the run.
- prec@i+: the number of questions with score i or above (i ∈ {2..4}) divided by the number of questions answered by the system. This measures the precision of the run, designed not to penalize non-answered questions.

Table 3 indicates the scores we obtained in the general domain question answering task. Our scores are close to the median. Yahoo seemed like a better knowledge source compared to Google for generating answers. One reason might be because questions are taken from Yahoo, and there is more

Table 3: Overall Ranking in General Domain Question Answering

Participant	avg	S@2	S@3	S@4	P@2	P@3	P@4
ECNU_ICA-EastChinaNormalUniversity	2.001	0.833	0.695	0.48	0.838	0.7	0.483
ECNU_ICA_2-EastChinaNormalUniversity	1.895	0.794	0.662	0.452	0.813	0.677	0.462
CMU-OAQA-CarnegieMellonUniversity	1.139	0.567	0.387	0.198	0.577	0.393	0.201
run2	0.777	0.421	0.25	0.126	0.482	0.286	0.144
run1	0.706	0.395	0.223	0.105	0.487	0.275	0.129
run3	0.613	0.357	0.19	0.084	0.434	0.231	0.102
ECNU-EastChinaNormalUniversity	0.206	0.14	0.059	0.016	0.141	0.059	0.016

Table 4: Over-ranking in medical sub task

	\boldsymbol{c}						
Participant	avg	S@2	S@3	S@4	P@2	P@3	P@4
CMU-OAQA-CarnegieMellonUniversity	0.637	0.392	0.265	0.098	0.404	0.273	0.101
run1	0.49	0.265	0.157	0.069	0.429	0.254	0.111
run2	0.441	0.275	0.137	0.059	0.394	0.197	0.085
run3	0.431	0.284	0.147	0.059	0.397	0.205	0.082
ECNU_ICA_2-EastChinaNormalUniversity	0.402	0.216	0.127	0.059	0.268	0.159	0.073
CMU-LiveMedQA-CarnegieMellonUniversity	0.353	0.216	0.137	0	0.218	0.139	0
ECNU_ICA-EastChinaNormalUniversity	0.255	0.225	0.147	0.029	0.228	0.149	0.03
ECNU-EastChinaNormalUniversity	0.137	0.216	0.088	0.01	0.216	0.088	0.01

overlap with the kind of topics discussed. The answer generation is dependent on the source of information and questions which are longer and have multiple sub-questions were hard to answer. Google⁴ search did not help much with the capture of candidate answers for the emotional and subjective questions.

Table 4 indicates the scores obtained for the medical sub-task. We used Wikipedia as our knowledge source, and as results show, Wikipedia has excellent coverage of medical concepts and diseases. Even though we have used only one system for medical subtask, the variability of the results is because of the default sources of knowledge used to answer general domain were tapped when the answer is not present in Wikipedia.

7 Conclusion

The performance of our open domain real-time question answering system is close to the median score of all runs in the competition. Not surprisingly, Yahoo answers website as our knowledge source performed well for general domain compared to the other knowledge sources we have used for other runs. For the medical sub-task, our system performed second-best among the runs in the competition. All the runs for the medical

sub-task had scores greater than the median score across all the runs in the competition. In the next steps, we would like to explore models for question understanding and emotional answer generation.

References

Vivek Datla, Sadid A Hasan, Ashequl Qadir, Kathy Lee, Yuan Ling, Joey Liu, and Oladimeji Farri. 2017. Automated clinical diagnosis: The role of content in various sections of a clinical document. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, pages 1004– 1011.

Vivek V Datla, Sadid A Hasan, Joey Liu, Yassine Benajiba, Kathy Lee, Ashequl Qadir, Aaditya Prakash, and Oladimeji Farri. 2016. Open domain real-time question answering based on semantic and syntactic question similarity. In TREC.

Sadid A Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, Oladimeji Farri, et al. 2016. Neural paraphrase generation with stacked residual lstm networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. pages 2923–2934.

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

⁴Google- https://google.com/

- S. Rose, D. Engel, N. Cramer, and W. Cowley. 2010. Automatic Keyword Extraction from Individual Documents. *Text Mining* pages 1–20.
- Abeed Sarker and Graciela Gonzalez. 2015. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics* 53:196–207.
- Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.