Powering COVID-19 community Q&A with Curated Side Information

Manisha Verma[1], Kapil Thadani[2] and Shaunak Mishra[3]

[1]*Yahoo! Research NYC*

[2]*Yahoo! Research NYC*

[3]*Yahoo! Research NYC*

Abstract

Community question answering and discussion platforms such as Reddit, Yahoo! answers or Quora provide users the flexibility of asking open ended questions to a large audience, and replies to such questions maybe useful both to the user and the community on certain topics such as health, sports or finance. Given the recent events around COVID-19, some of these platforms have attracted 2000+ questions from users about several aspects associated with the disease. Given the impact of this disease on general public, in this work we investigate ways to improve the ranking of user generated answers on COVID-19. We specifically explore the utility of external technical sources of side information (such as CDC guidelines or WHO FAQs) in improving answer ranking on such platforms. We found that ranking user answers based on question-answer similarity is not sufficient, and existing models cannot effectively exploit external (side) information. In this work, we demonstrate the effectiveness of different attention based neural models that can directly exploit side information available in technical documents or verified forums (e.g., research publications on COVID-19 or WHO website). Augmented with a temperature mechanism, the attention based neural models can selectively determine the relevance of side information for a given user question, while ranking answers.

Keywords

question answering, deep learning, knowledge injection, NLP

1. Introduction

Question answering systems are key to finding relevant and timely information about several issues. Community question answering (cQ&A) platforms such as Reddit, Yahoo! answers or Quora have been used to ask questions about wide ranging topics. Most of these platforms let users ask, answer, vote or comment on questions present on the platform. However, question answering platforms are useful not only for getting public opinions or votes about areas such as entertainment or sports but can also serve as information hot-spots for more sensitive topics such as health, injuries or legal topics. Thus, it is imperative that when the user visits *sensitive* topics content, answer ranking also takes into account curated side information from reliable (external) sources. Most prior work on cQ&A has focused on incorporating question-answer similarity [1, 2], user reputation [3, 4, 5], integration of multi-modal content [6], community interaction features [7] associated with answers or just the question answering network [8] on the platform. However, there is very limited work on incorporating *curated content* from external sources. Existing work only exploits knowledge bases [6] that consist of different entities and relationships be-

tween these entities to score answers. However, there are some limitations of knowledge bases that would make it difficult to use them for community Q&A for rapidly evolving topics such as disease outbreaks (e.g. ebola, COVID-19), wild-fires or earthquakes. Firstly, knowledge bases contain information about established entities, and do not rapidly evolve to incorporate new information which makes them unreliable for novel disease outbreaks such as COVID-19 where information rapidly changes and its verification is time sensitive. Secondly, it may be hard to determine what even *constitutes an entity* as new information arrives about the topic. To overcome these limitation in this work, we posit that external *curated* free-text or semi-structured informational sources can also be used effectively for cQ&A tasks.

In this work, we demonstrate that free text or semi structured external information sources such as CDC[1], WHO[2] or NHS[3] can be very useful for ranking answers on community Q&A platforms since they contain frequently updated information about several topics such as ongoing disease outbreaks, vaccines or resources about other topics such as surgeries, birth control or historical numerical data about diseases across the world.

We argue that for sensitive topics such as COVID-19, it is useful to use publicly available *vetted* information for improving our ranking systems. In this work, we explore

[1]https://www.cdc.gov/
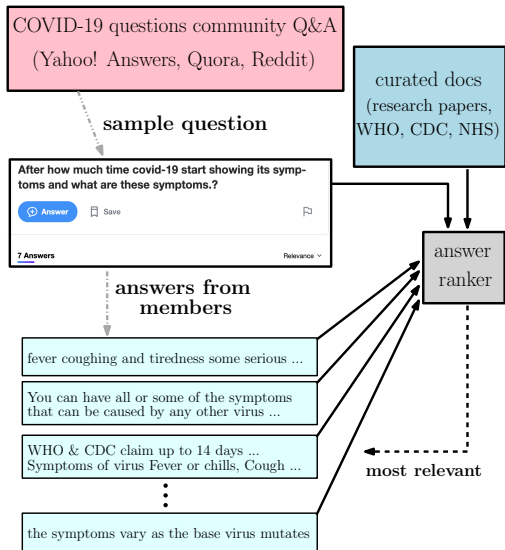[2]https://www.who.int/
[3]https://www.nhs.uk/

Figure 1: Illustrative example of COVID-19 community answer ranking powered by side information in the form of research papers, and information from verified sources (such as CDC, WHO, and NHS).

the utility of publicly available information for ranking answers for questions associated with COVID-19. We specifically focus on ranking answers for questions in two publicly available *primary* Q&A datasets: a) Yahoo! Answers[4] and b) recently released annotated Q&A dataset [9] in presence of two *external* semi-structured curated sources: a) TREC-COVID [10] and b) WHO questions and answers [5] on COVID-19. We explore the utility of deep learning models with attention in this work to improve upon existing state-of-the-art systems.

More specifically, we propose a temperature regulated attention mechanism to rank answers in presence of external (side) information. Our experiments on 10K+ questions from both *source* datasets on COVID-19 show that our models can improve ranking quality by a significant margin over question-answer matching baselines in presence of external information. Figure 1 demonstrates the overall design of our system. We specifically use attention based neural architecture with temperature to automatically determine which components in the external information are useful for ranking user answers with respect to a question. Ranking performance, when evaluated with three metrics shows that precision and recall for correct answer retrieval improves by ∼17% and ∼9% for both source datasets respectively over several other cQ&A models.

2. Related work

Community Question and Answering (cQ&A) systems is a well researched sub-field both in information retrieval and NLP communities. Several systems have been proposed to rank user submitted answers to questions on community platforms such as Yahoo! answers, Reddit and Quora.

Ranking user submitted answers on community question-answering platforms has been addressed with several approaches. Primary method is to determine the relevance of the answer given an input question. Text based matching is one of the most common approaches to rank answers. Researchers have used several methods to compute *similarity* between a question and user generated answers to determine relevance. For instance, feature based question-answer matching is used in [1] with 17 features extracted from unigrams, bigrams and web correlation features using unstructured user search logs to rank answers. It is worth noting that user features and community features when incorporated may still yield further improvements in the performance of these models but this is not the focus of our work. The authors in [1] used questions extracted from Yahoo! answers for their experiments. Researchers have used different approaches such representation learning, for instance, in [11, 12] authors use LSTM to represent questions and answers respectively. Convolutional networks have also been used in [3, 13] to rank answers. Other approaches such as doc2vec [14], tree-kernels [15], adversarial learning [16], attention [17, 6, 11, 18] or deep belief networks [19] have been used to score question and answer pairs. There have also been studies exploring community, user interaction or question based features [3, 4, 5, 7] to rank answers. While these approaches are relevant, it is not always evident how one can incorporate external information when it is either in free-text or semi-structured format into these systems. We explore some question-answer based matching approaches as baselines in this work and show that for rapidly evolving topics such as COVID-19, inclusion of external *curated* information can boost model performance.

The line of work most closely related to ours is incorporation of knowledge bases in Q&A systems. Existing work [20, 21, 6], however, approaches different tasks. For instance, authors in [21, 20] focus on finding factual answers to questions using a knowledge base. This does not extend easily to cQ&A where neither the questions nor the answers may request or refer to any facts. Most recent work is [6] on incorporating medical KB for ranking answers on medical Q&A platforms. They propose to learn path based representation of *entities (from KB)* present in question and answers posted by users. This approach relies on reliable detection of entities first, which may be absent for emerging topics such as COVID-19

pandemic. Another limitation of this work is that external knowledge may not always be present in a *structured* format. For example, CDC guidelines are usually simple question-answer pairs posted on the website. This makes it difficult to apply their approach to our problem. The proposed approach in this work incorporates semi-structured information directly with help of temperature regulated attention.

Finally, with the rise of COVID-19, researchers across disciplines are actively publishing information and datasets to share understanding of the virus and its impact on people. Researchers routinely organize dedicated challenges such as SemEval [22] with tasks such as ranking answers on QA forums. One such initiative is TREC-COVID track [10] which released queries, documents and manual relevance judgements to power search for COVID related information. Authors in [23] also released COVID-19 related QA dataset with 100+ questions and answers pairs extracted from TREC COVID [6] initiative. These questions/answer pairs are not user generated content, hence, do not reflect real user questions. We also rely on recently released Q&A dataset from [9] for our task. We also compile a dataset of 2000+ COVID-19 questions with 10K+ answers all submitted by users on Yahoo! answers for this work.

3. Method

3.1. Problem formulation

In this work, we focus on ranking answers for n questions $q_1, \ldots, q_n$ related to an emerging topics such as COVID. Each q_i is associated with a set of two or more answers $A_i = \{a_{ij} : j \geq 2\}$ and corresponding labels $Y_i = \{y_{ij} : j \geq 2\}$ representing answer relevance. We use binary indicator for relevance where relevant judgments (e.g., *favorite*, *upvoted*) are provided by the user, i.e., $y_{ij} \in \{0, 1\}$ respectively.

We attempt to model the relevance of each answer a_{ij} to its corresponding question using an external source which may contain free text or semi-structured information. For example, the *external* source could consist of information-seeking queries or questions $eq_1, \ldots, eq_m$ related to a topic, with each eq_k linked to a set of relevant scientific articles or answers ED_k, where each answer/document $ed_1, \ldots, ed_p$ may be judged for relevance by human judges [10] or some experts.

We hypothesize that this semi-structured or free-text information may be valuable in identifying user answer quality for certain kinds of questions, although not all. We investigate this with our model to recover the true labels y_{ij} for each user answer $a_{ij} \in A_i$ given its question q_i, category information, and information from the
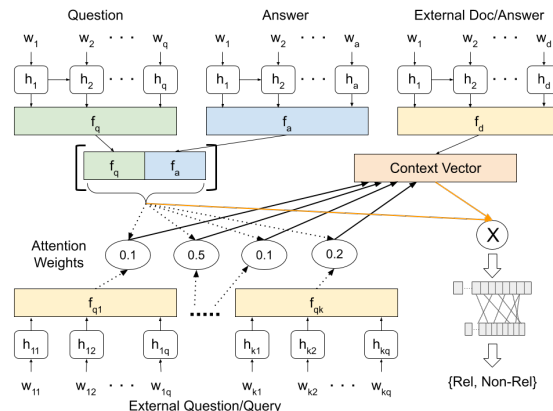

Figure 2: External source augmentation model

external source $\langle eq_k, ED_k \rangle_{k=1}^m$.

3.2. Proposed Model

In this work, we explore token-level matching mechanism to determine the relevance of information in the external source that may inform the label prediction task. Our model (τ-*att*) aims to match a given user question with all the submitted answers in the presence of external information about the same domain. First, the question q_i, an answer a_{ij} and additional metadata can be encoded into a d-dimensional vector x_i using a text encoder f_{input}. We use LSTM based encoder for both question and answer in the *primary* source which can handle input sequences of variable length.

Question Encoding: Each word w_i^q in a question is represented as a K dimensional vector with pre-trained word embeddings. LSTM takes each token embedding as input and updates hidden state h_i^q based on previous state h_{i-1}^q. Finally, the hidden state is input to a feed forward layer with smaller dimension $F < K$ to compress question encoding as follows:

$$h_i^q = LSTM(h_{i-1}^q, w_i^q), \ f_i^q = RELU(h_i^q W_q + b_q) \tag{1}$$

Answer Encoding: Each word w_j^a in the answer is also represented as a K dimensional vector with pre-trained word embeddings. LSTM takes each token embedding as input and updates hidden state h_j^a. We also reduce the dimension of answer encoding with a feed forward layer with dimension $F < K$ as follows:

$$h_j^a = LSTM(h_{j-1}^a, w_j^a), \ f_j^a = RELU(h_j^a W_a + b_a) \tag{2}$$

Source	Question	Rel answer	Non-rel answer
Yahoo! Ans	I am really scared to go places for St. Patrick's day because of the coronavirus. what do I do?	Unfortunately, there's not enough people that care and will still go out and party despite the coronavirus epidemic. I'm proud of you in that you're taking extra precautions ... Good for you!	Stop being scared of viruses. What's the problem?
Infobot	Can corona live on cardboard?	A recent study shows that the virus can live in the air ... On cardboard, it can live up to 24 hours (1 day)	The risk is quite low for one to become infected with COVID19 through mail/packages - especially because...(over a period of a few days/weeks).

Table 1
Sample rel/non-rel answers from both sources

We concatenate the question and answer representations for further processing.

$$f_{ij} = [f_i^q, f_j^a] \tag{3}$$

External source encoding: External sources of information can vary from task-to-task. We encode each segment of data individually. For instance, if there are two segments in the source (e.g. question/answer or query/document), our system encodes both segments individually. We use the same encoding architecture used for primary source question/answer encoding above. Encoding example for two segment *external* source is given below.

$$h_t^{eq} = LSTM(h_{t-1}^{eq}, w_t^{eq}), \ f_t^{eq} = RELU(h_t^{eq}W_{eq} + b_{eq})$$
$$h_t^{ed} = LSTM(h_{t-1}^{ed}, w_t^{ed}), \ f_t^{ed} = RELU(h_t^{ed}W_{ed} + b_{ed}) \tag{4}$$

We incorporate external source encoding with a temperature (τ) based variant of scaled dot-product attention, which provides a straightforward conditioning approach over a set of query-document pairs. Question encoding vector f_{ij} serves as a query over keys f_t^{eq}. If two segments are present in the external source such as query/document, the model uses the attention weights over first segment (e.g. query) to determine the importance of the second segment (e.g. document) respectively. It is easy to extend this framework to *external* sources with multiple segments. The two segment attention is described below.

$$z_{it} = \frac{f_{ij}^\top f_t^{eq}}{\sqrt{d}}$$
$$\alpha_{it} = \frac{e^{z_{it}/\tau}}{\sum_l e^{z_l/\tau}} \tag{5}$$
$$s'_{itd} = \sum_d \alpha_{it} f_t^{ed}$$

To summarize, temperature (τ) based attention helps determine the relevance of each f_t^{ed} corresponding f_t^{eq}

with respect to the question encoding. Temperature (τ) parameter helps us control the uniformity of attention weights α_{it}. Finally, labels are predicted using a multi-layer perceptron over the input vector f_{ij} and the learned weighted average of side information s'_{itd}. We use binary cross entropy loss to train the proposed model.

$$\hat{y}_{ij} = F_{\text{output}}([f_{ij}; s'_{itd}])$$

where F_{output} uses sigmoid activation function. Since community questions may often be entirely unrelated to external sources, a key aspect of this approach is determining *whether* the *external source* is useful, not merely attending to its entries that are most relevant. Temperature based attention mechanism is useful in controlling which external source entries are useful for user questions. It is worth noting that one will have to experiment and tune the value of temperature τ such that ranking performance improves.

4. Experimental Setup

Given the model architecture, in this section, we provide a detailed overview of different datasets, metrics and baselines used in our experiments.

4.1. Data

We compiled two question answering datasets. The first was collected from Yahoo! answers and the second was recently released in [9] where both datasets have questions raised by real users. In this work we focus specifically on questions associated with COVID-19. Different statistics about the train and test split of both q&a datasets are given in Table 2 respectively. A pair of relevant and non-relevant answers for a question in both datasets is also shown in Table 1 for reference. More details about them is given below.
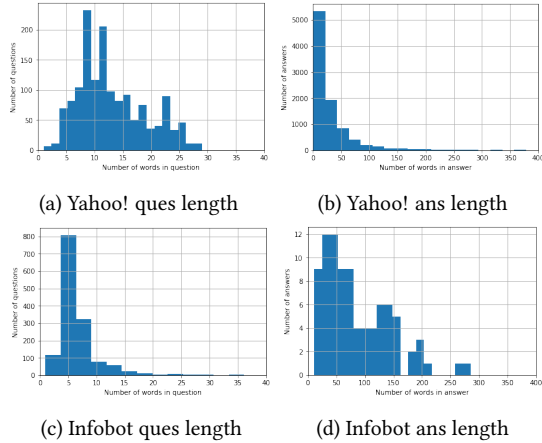
(a) Yahoo! ques length   (b) Yahoo! ans length

(c) Infobot ques length   (d) Infobot ans length

Figure 3: Token distribution in different sources

Stat	Yahoo! Ans	Infobot
Train Q-A	9341	6354
Train ans/q	6.25±2.9	4.40±0.77
Train #qwords	12.71±5.8	6.55±3.93
Train #awords	36.31±93.59	92.17±59.27
Test Q-A	2232	1592
Test ans/q	5.96±2.87	4.41±0.76
Test #qwords	13.07±5.89	6.21±2.94
Test #awords	35.64±80.31	92.39±59.47

Table 2
Train and test data from primary sources

Yahoo! Dataset : We crawled COVID-19 related questions from Yahoo! answers [7] using several keywords such as 'coronavirus', 'covid-19', 'covid', 'sars-cov2' and 'corona virus' between the period of Jan 2020 to July 2020 to ensure we gather all possible questions for our experiments. We keep only those questions have two or more answers. In total, we obtained 1880 questions with 11500 answers. We used favorite answers as positive labels (similar to previous work [1]), assuming that users, over time rate answers (with upvotes/downvotes) that are most relevant to the submitted question. We normalized the question and answer text by removing a small list of stop words, numbers, links or any symbols. Figure 3a and 3b show the distribution of question and answer lengths respectively. Questions contain 12.7 ± 5.8 *(qwords)* words and answers consist of 36.3 ± 93.5 (mean±std) words *(awords)* respectively which indicates that user submitted answers can vary widely on Yahoo! answers. On average, a question has about 6 answers *(ans/q)* in Yahoo! ans dataset. We spilt the data into three sets: train (64%, 1196 questions, 7435 answers), validation (16%, 298 questions, 1858 answers) and test (20%, 374 questions, 2310 answers)

set where questions for each set were uniformly sampled.

Infobot Dataset [9] : Researchers at JHU [9] have recently compiled a list of user submitted questions on different platforms and manually labeled 22K+ question-answer pairs. We cleaned this set by removing questions with less than two answers or no relevant answers. In total, our dataset contains 8000+ question answer pairs where each question may have *multiple* relevant answers which is not the same as Yahoo! answers dataset. Figure 3c and 3d show the distribution of question and answer lengths respectively.

4.1.1. External sources

We use two external datasets to rank answers. Details of each dataset are given below:

TREC COVID [10]: We use recently released TREC COVID-19 track data with 50 queries which also contain manually drafted query descriptions and narratives. Expert judges have labeled over 5000 scientific documents for these 50 queries from the CORD-19 dataset [8]. These documents contain coronavirus related research. Given the documents are scientific literature, we initialize document embeddings using SPECTER [24].

WHO: We use data released on question and answer hub of WHO[9] website to create a list of question-answer pairs. There are 147 question and answer pairs in this dataset where questions contain 13.28±5.36 words and answers contain 133.2±100.9 words respectively.

4.2. Baselines

We evaluated our model against embedding similarity baseline. We computed four baselines as follows:

Random: An answer is chosen at random as relevant for a user question. This is expected to provide a lower bound on retrieval performance.

Linear Attention (*att*) : When $\tau = 1.0$, our model defaults to simple linear attention over all the information present in the external sources. This gives an indication of how well the model performs when its forced to look at all the information in the external source.

---

[7]https://answers.search.yahoo.com/search?p=coronavirus

[8]https://www.semanticscholar.org/cord19
[9]https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub

Model	Yahoo! Ans			Infobot		
	P@1	R@3	MRR	P@1	R@3	MRR
τ-att	0.393	0.644	0.598	0.673	0.868	0.802
λ-sim	0.3743	0.633	0.578	0.551	0.817	0.7207
bsl-256	0.406	0.657	0.615	0.581	0.803	0.744
bsl-128	0.363	0.604	0.589	0.557	0.799	0.731
att	0.377	0.645	0.589	0.567	0.821	0.739
qasim	0.318	0.608	0.546	0.551	0.817	0.720
random	0.21	-	-	0.239	-	-

Table 3
Evaluation with WHO external data

Linear combination (λ-*sim*) : We linearly combine similarities between Yahoo! question-answer and Trec query-answer as shown below:

$$\lambda\text{-}sim = \lambda \, cos(ya, yq) + (1 - \lambda) \, \max_{tq}(cos(ya, tq)) \quad (6)$$

where ya, yq and tq are Yahoo! answer, question and concatenated trec query, narrative and description embeddings respectively. This is a more crude version of temperature attention where λ controls the contribution of each component directly. We vary λ to determine the optimal combination. Question-Answer similarity (*qasim*) is similarity between question and answer embedding i.e. $\lambda = 1$. Both question and answer embeddings are obtained by averaging over their individual token embeddings.

BERT Q&A (*bert*) : Large scale pre-trained transformers [25] are widely popular for NLP tasks. BERT like models have shown effectiveness on Q&A datasets such as SQUAD [10]. We fine-tune BERT base model with two different answer lengths a) 128 *(bsl-128)* and b) 256 tokens *(bsl-256)* respectively. The intuition is that large scale pre-trained models are adept at language understanding and can be fine-tuned for new tasks with small number of samples. We finetune BERT for both datasets Yahoo! ans and Infobot respectively. It is non-trivial to include external information in BERT and we leave this for future work.

4.3. Evaluation Metrics

We evaluate the performance of our model using three popular ranking metrics, mainly Precision (P@1), Mean Reciprocal Rank (MRR), and Recall (R@3). Each metric is described below:

- **Precision (P@k):** Precision at position k evaluates the fraction of relevant answers retrieved until position k. For, both datasets Yahoo! ans and Infobot [9], we evaluate whether the top answer

i.e. ($k = 1$) in the ranked list is indeed correct. It is defined as follows:

$$Prec@k = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{\sum_{j=1}^{k} \mathbb{I}\{rel_{ij} = 1\}}{k} \quad (7)$$

where $\mathbb{I}\{rel_{ij} = 1\}$ indicates whether the answer at position j is relevant to the ith question.

- **Recall (R@k):** Recall at position k evaluates the fraction of relevant answers retrieved from all the answers marked relevant for a question. We report recall averaged for all the queries in test set. For recall, we take a cutoff as ($k = 3$), which evaluates whether the model is able to retrieve the correct answers in top 3 positions. It is defined as follows:

$$Recall@k = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{\sum_{j=1}^{k} \mathbb{I}\{rel_{ij} = 1\}}{|rel_i|} \quad (8)$$

where $|rel_i|$ is the number of relevant answers for the ith question.

- **MRR (MRR):** evaluates the average of the reciprocal ranks corresponding to the most relevant answer for the questions in test set, which is given by:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \quad (9)$$

where $|Q|$ indicates the number of queries in the test set and $rank_i$ is the rank of the *first* relevant answer for the ith query.

4.4. Parameter Settings

Both primary datasets, Yahoo! ans and Infobot, were divied into three parts: train ($\sim$60%), validation and test (20%) respectively. The baseline models λ-sim and att are initialized with glove embeddings [11] of 100 dimensions. We performed a parameter sweep over λ and τ for λ-sim and τ-att models with step size of 0.1 between $\{0, 1.0\}$ respectively. We used base uncased model for $bert$ implementation. We fine-tuned the model between 1-10 epochs and found that 3 epochs gave the best result on validation set. We used LSTM with 64 hidden units to represent question, answer and all the information in external datasets. We experimented with higher embedding size and hidden units, but the performance degraded significantly as the model tends to overfit on training data. Lastly we used batch size of 64 and trained the model for 30 epochs with early stopping.

---

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

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

	Yahoo! Ans			Infobot		
Model	P@1	R@3	MRR	P@1	R@3	MRR
τ-att	0.532	0.778	0.715	0.606	0.842	0.766
λ-sim	0.326	0.616	0.555	0.556	0.813	0.722
bsl-256	0.406	0.657	0.615	0.581	0.803	0.744
bsl-128	0.363	0.604	0.589	0.557	0.799	0.731
att	0.291	0.495	0.494	0.601	0.833	0.762
qasim	0.318	0.608	0.546	0.551	0.817	0.720
random	0.21	-	-	0.239	-	-

Table 4
Evaluation with TREC-COVID external data

Category	τ-att	λ-sim	qasim
Entertainment (47)	0.829	0.702	0.59
Health (62)	0.693	0.69	0.645
Politics (143)	0.727	0.629	0.587
Society (38)	0.578	0.473	0.42
Family (20)	0.85	0.750	0.65

Table 5
Recall@3 of models across categories

Category	τ-att	λ-sim	qasim
Entertainment (47)	0.446	0.382	0.297
Health (62)	0.483	0.419	0.354
Politics (143)	0.45	0.300	0.272
Society (38)	0.28	0.157	0.236
Family (20)	0.6	0.350	0.40

Table 6
Precision@1 of models across categories

5. Results

In this work, our focus is to evaluate the utility of external information in improving answer ranking for cQ&A task. Thus, we performed experiments to answer three main research questions listed below.

RQ1: Does external information improve answer ranking?

RQ2: How does temperature (τ) compare with λ parameter?

RQ3: What kind of queries/questions does the model attend to when ranking relevant/non-relevant answers?

RQ1: Does external information improve answer ranking? We evaluated different models for ranking answers in Yahoo! ans and Infobot dataset in presence of TREC and WHO datasets respectively. We found that temperature regulated attention models that incorporate external sources indeed outperform the baselines as shown in Table 4 and Table 3 respectively. (τ-att) model beats *bert* models by ∼30% in precision, ∼18% in recall and ∼16% in MRR respectively on TREC data. However, (τ-att) does only marginally better than *att* model in precision and MRR on Infobot data. We suspect that is due to the large set of query-document pairs in TREC-COVID data compared to fewer number of question-answer pairs in Infobot dataset. Our results also clearly suggest that embedding based matching of question-answer pair (*qasim*) would not yield a good ranker, though it is better than choosing an answer at random (*random*). When WHO is used as an external dataset, we find that (τ-att) model is slightly worse than *bert*. This suggests that not all sources would equally ben-

efit cQ&A task. Since attention is dependent on the input query and key embedding lengths, it would be interesting to scale the computation in our model to incorporate several open external datasets to overcome this limitation in the future.

Yahoo! ans questions are also assigned categories by users. Category based breakdown of performance on test set is given in Table 6 and Table 5 respectively, where categories with largest number of questions in test set are listed. In all the categories, our model outperforms best λ-sim and *qasim* model respectively. The largest improvement happens for questions in Family category where our model achieves an improvement of 71% over the λ-sim model. It seems that ranking answers for questions from society and politics are harder than other categories. All the models, however, are able to rank the top answer in first three positions effectively as Recall@3 is high for all the categories.

RQ2: How does temperature (τ) compare with λ parameter? We argued that linearly combining similarities between question-answer in primary dataset and between question-external source may not be sufficient to boost performance. We observe that in our results too i.e. λ-sim models do not perform better than (τ-att) models. This clearly indicates that more sophisticated models can learn to combine this information directly from training data. However, our experiments indicate that optimal value of (τ) varies across primary datasets and external sources of information. For instance, (τ-att) model performed best when $\tau = 0.4$ and $\tau = 0.9$ for Yahoo! ans and Infobot dataset respectively when TREC was used as external source. It performed best when $\tau = 0.1$ and $\tau = 0.5$ for Yahoo! ans and Infobot dataset respectively when WHO was used as external source. We also tried to vary τ beyond 1.0 to determine whether it yielded a trend as shown in Table 7. Higher values of temperature seem to degrade model performance. We found that optimal temperature range is between $[0.1 - 1]$. Existing research in model distillation [26] has also empirically found that lower values of temperature yield better performance.

We also compared model performance in terms of precision when λ and τ are varied for λ-sim models and temperature based models respectively as shown in Figure 4. Temperature based models peak at one value but
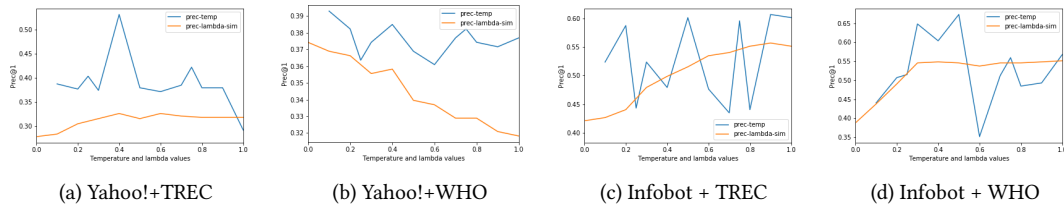
Figure 4: Temperature and λ variation impact on Prec@1

	Temperature (τ) > 1.0					
	10	100	1000	10	100	1000
Src+ Ext	Prec@1			Recall@3		
Y! + TREC	0.46	0.38	0.38	0.73	0.644	0.64
Y! + WHO	0.37	0.38	0.36	0.64	0.65	0.64
Ibot + TREC	0.44	0.59	0.39	0.72	0.81	0.75
Ibot + WHO	0.65	0.41	0.44	0.85	0.76	0.79

Table 7
Variation in P@1 and R@3 across different temperature values.

Figure 5: Y! ques, its rel and non-rel ans and questions with τ-att model's attention values for TREC queries.

Figure 6: Infobot ques, its rel and non-rel ans and questions with τ-att model's attention values for TREC queries.

do not have a clear trend indicating that one needs to explore different τ values at the time of training for better performance. On the other hand, we observe that adding external information also helps the λ-sim models until a certain threshold. Overall, both sets of models show that free-text external information can be incorporated to improve answer ranking performance.

RQ3: What kind of queries/questions does the model attend to when ranking relevant/non-relevant answers? Attention based models have a very unique feature: they can aid explaining the internal workings of neural network models. We inspect what kind of queries/questions in external datasets does our model pay attention to while ranking relevant or non-relevant answers. Figure 5 shows one such example of Yahoo! question and incorporation of TREC data. At the time of scoring relevant answer, the model gives higher weight to some queries compared to others. In the example, for instance, it assigns more weight to queries associated with masks or COVID virus response to weather changes. We observe higher attention weights for questions when relevant answers are ranked than when non-relevant answers are scored. An example question, a relevant and non-relevant answer along with model attention weights on TREC queries are shown from the Infobot data in Figure 6 respectively. It shows a similar trend where attention weights are high for external queries that are closely associated with the question answer text.

Overall, our experiments show that curated external information is useful for improving community question answering task. Our experiments also indicate that
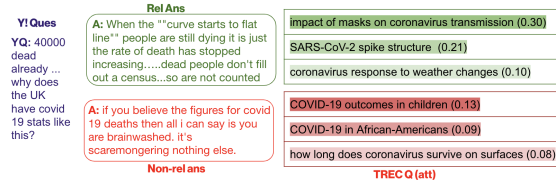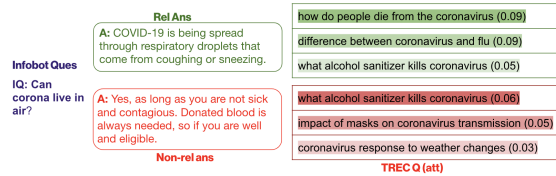
this external knowledge need not always be structured text. However, it is worth noting that curated and reliable external sources may not always be available for all domains. We addressed a very niche task in this work, and further research is required to extend it to incorporate multiple external sources. We posit that with scalable attention mechanisms, this work can be easily made tractable for large external sources containing thousands or millions of entries in the future.

6. Conclusion

Question answering platforms provide users with effective and easy access to information. These platforms also provide content on rapidly evolving *sensitive* topics such as disease outbreaks (such as COVID-19) where it is also useful to use external *vetted* information for ranking answers. Existing work only exploits knowledge bases which have some limitations that makes it difficult to use them for community Q&A for rapidly evolving topics such as wild-fires or earthquakes. In this work, we tried to evaluate the effectiveness of external (free text or

semi-structured) information in improving answer ranking models. We argue that simple question-answer text matching may be insufficient and in presence of external knowledge, but temperature regulated attention models can distill information better which in turn yields higher performance. Our proposed model with temperature regulated attention, when evaluated on two public datasets showed significant improvements by augmenting information from two *external* curated sources of information. In future, we aim to expand these experiments to other categories such as disaster relief and scale the attention mechanism to include multiple external sources in one model.

References

[1] M. Surdeanu, M. Ciaramita, H. Zaragoza, Learning to rank answers on large online qa collections, in: ACL, 2008.

[2] Y. Shen, W. Rong, Z. Sun, Y. Ouyang, Z. Xiong, Question/answer matching for cqa system via combining lexical and sequential information, in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI'15, AAAI Press, 2015, p. 275–281.

[3] L. Yang, Q. Ai, D. Spina, R.-C. Chen, L. Pang, W. B. Croft, J. Guo, F. Scholer, Beyond factoid qa: Effective methods for non-factoid answer sentence retrieval, in: European Conference on Information Retrieval, Springer, 2016, pp. 115–128.

[4] L. Hong, B. D. Davison, A classification-based approach to question answering in discussion boards, in: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009, pp. 171–178.

[5] D. H. Dalip, M. A. Gonçalves, M. Cristo, P. Calado, Exploiting user feedback to learn to rank answers in qa forums: A case study with stack overflow, in: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 543–552. URL: https://doi.org/10.1145/2484028.2484072. doi:10.1145/2484028.2484072.

[6] Y. Zhang, S. Qian, Q. Fang, C. Xu, Multi-modal knowledge-aware hierarchical attention network for explainable medical question answering, in: Proceedings of the 27th ACM International Conference on Multimedia, MM '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 1089–1097. URL: https://doi.org/10.1145/3343031.3351033. doi:10.1145/3343031.3351033.

[7] J. Hu, S. Qian, Q. Fang, C. Xu, Attentive interactive convolutional matching for community question answering in social multimedia, in: Proceedings of the 26th ACM International Conference on Multimedia, MM '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 456–464. URL: https://doi.org/10.1145/3240508.3240626. doi:10.1145/3240508.3240626.

[8] J. Hu, S. Qian, Q. Fang, C. Xu, Hierarchical graph semantic pooling network for multi-modal community question answer matching, in: Proceedings of the 27th ACM International Conference on Multimedia, MM '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 1157–1165. URL: https://doi.org/10.1145/3343031.3350966. doi:10.1145/3343031.3350966.

[9] A. Poliak, M. Fleming, C. Costello, K. W. Murray, M. Yarmohammadi, S. Pandya, D. Irani, M. Agarwal, U. Sharma, S. Sun, et al., Collecting verified covid-19 question answer pairs (2020).

[10] E. Voorhees, T. Alam, S. Bedrick, D. Demner-Fushman, W. R. Hersh, K. Lo, K. Roberts, I. Soboroff, L. L. Wang, Trec-covid: Constructing a pandemic information retrieval test collection, 2020. arXiv:2005.04474.

[11] A. Rücklé, I. Gurevych, Representation learning for answer selection with LSTM-based importance weighting, in: IWCS 2017 — 12th International Conference on Computational Semantics — Short papers, 2017. URL: https://www.aclweb.org/anthology/W17-6935.

[12] D. Cohen, W. Croft, End to end long short term memory networks for non-factoid question answering, 2016, pp. 143–146. doi:10.1145/2970398.2970438.

[13] X. Zhou, B. Hu, Q. Chen, B. Tang, X. Wang, Answer sequence learning with neural networks for answer selection in community question answering, arXiv preprint arXiv:1506.06490 (2015).

[14] L. Nie, X. Wei, D. Zhang, X. Wang, Z. Gao, Y. Yang, Data-driven answer selection in community qa systems, IEEE transactions on knowledge and data engineering 29 (2017) 1186–1198.

[15] A. Severyn, A. Moschitti, Structural relationships for large-scale learning of answer re-ranking, in: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 2012, pp. 741–750.

[16] X. Yang, M. Khabsa, M. Wang, W. Wang, A. H. Awadallah, D. Kifer, C. L. Giles, Adversarial training for community question answer selection based on multi-scale matching, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 395–402.

[17] H. Huang, X. Wei, L. Nie, X. Mao, X.-S. Xu, From question to text: Question-oriented feature attention for answer selection, ACM Transactions on

Information Systems 37 (2018) 1–33. doi:10.1145/3233771.

[18] X. Zhang, S. Li, L. Sha, H. Wang, Attentive interactive neural networks for answer selection in community question answering, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, AAAI Press, 2017, p. 3525–3531.

[19] B. Wang, X. Wang, C.-J. Sun, B. Liu, L. Sun, Modeling semantic relevance for question-answer pairs in web social communities, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 1230–1238.

[20] B. Kratzwald, A. Eigenmann, S. Feuerriegel, Rankqa: Neural question answering with answer re-ranking, CoRR abs/1906.03008 (2019). URL: http://arxiv.org/abs/1906.03008. arXiv:1906.03008.

[21] Y. Shen, Y. Deng, M. Yang, Y. Li, N. Du, W. Fan, K. Lei, Knowledge-aware attentive neural network for ranking question answer pairs, in: The 41st International ACM SIGIR Conference on Research Development in Information Retrieval, SIGIR '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 901–904. URL: https://doi.org/10.1145/3209978.3210081. doi:10.1145/3209978.3210081.

[22] P. Nakov, D. Hoogeveen, L. Màrquez, A. Moschitti, H. Mubarak, T. Baldwin, K. Verspoor, Semeval-2017 task 3: Community question answering, arXiv preprint arXiv:1912.00730 (2019).

[23] D. Su, Y. Xu, T. Yu, F. B. Siddique, E. J. Barezi, P. Fung, CAiRE-COVID: A question answering and multi-document summarization system for covid-19 research, arXiv 2005.03975 (2020).

[24] A. Cohan, S. Feldman, I. Beltagy, D. Downey, D. S. Weld, Specter: Document-level representation learning using citation-informed transformers, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2270–2282.

[25] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[26] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, 2015. arXiv:1503.02531.