Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information (Supplementary Material)

Sudha Rao
University of Maryland, College Park
raosudha@cs.umd.edu

Hal Daumé III
University of Maryland, College Park
Microsoft Research, New York City
hal@cs.umd.edu

1 Introduction

In this supplementary material, we add additional details supporting the dataset §2, experiments §3 and expert annotation process §4 described in the main paper. We also show the expert annotations on three example posts and compare it to our EVPI model predictions.

2 Data analysis

How often are extracted questions clarifications? A natural question to our process of data creation would be how often is the extracted question a clarification question. We sample a set of 1000 questions from our dataset and design a crowdsourced task on CrowdFlower[1] where given a question we ask annotators to choose whether the question was: (a) Asking for more information, (b) Providing an answer or a suggestion; or (c) Neither. We collect three annotations per question. We find that 91% of the questions were marked with option (a), 7% with option (b) and 2% with option (c). These numbers suggest that a large portion of the extracted questions are indeed "clarification questions". Additionally, we analyze the questions marked as "providing a solution" and find that majority of these started with one of the following phrases: *"have you", "did you try", "can you try", "could you try"*. We preprocess our dataset to remove all such instances.

How useful are clarifications questions? A clarification question is useful if it helps in generating an answer for a given post. Imagine a scenario in which a post goes unanswered for sometime. Following this, a clarification question gets asked on this post and then the post gets an answer. Such a scenario will help showcase the usefulness of clarification questions. We estimate such a usefulness by calculating the following two probabilities for posts that have not received an answer within a week:

$$Pr(A|CQ) = \frac{\#(A|CQ)}{\#(A|CQ) + \#(\neg A|CQ)}$$

$$Pr(A|\neg CQ) = \frac{\#(A|\neg CQ)}{\#(A|\neg CQ) + \#(\neg A|\neg CQ)}$$

where:

$\#(A|CQ)$: # answered posts with a clarification question

$\#(\neg A|CQ)$: # unanswered posts with a clarification question

$\#(A|\neg CQ)$: # answered posts without a clarification question

$\#(\neg A|\neg CQ)$: # unanswered posts without a clarification question

	askubuntu	unix	superuser
$Pr(A\|CQ)$	0.82	0.85	0.45
$Pr(A\|\neg CQ)$	0.77	0.80	0.34

Table 1: Likelihood of a post getting answered with and without a clarification question

Table 1 shows these probabilities for the three data domains. We can see that, overall, the likelihood of a post getting an answer with a clarification question is higher than the the likelihood of a post getting an answer without a clarification question.

Yes/No clarification questions. We argued in the introduction of the main paper that asking a question like "What version of Ubuntu do you have?" is more useful than asking a more specific question that might yield a Yes/No answer. This raises the question of how many clarification questions in our dataset are Yes/No questions. We manually inspect 100 randomly selected clarification questions in our dataset and find that 13 of them were Yes/No questions. This suggests that users, on these forums, tend to ask questions that are generic enough to elicit a useful answer more than a specific question.

[1] www.crowdflower.com

Model	$B1 \cup B2$ p@1	p@3	p@5	MAP	$V1 \cap V2$ p@1	p@3	p@5	MAP
Random	17.4	17.5	17.5	26.7	26.3	26.4	26.4	37.0
Bag-of-ngrams	16.3	18.9	17.5	25.2	26.7	28.3	26.8	37.3
Community QA	22.6	20.6	18.6	29.3	30.2	29.4	27.4	38.5
Neural (p,q)	20.6	20.1	18.7	27.8	29.0	29.0	27.8	38.9
Neural (p,a)	22.6	20.1	18.3	28.9	30.5	28.6	26.3	37.9
Neural (p,q,a)	22.2	21.1	19.9	28.5	29.7	29.7	28.0	38.7
EVPI	23.7	21.2	19.4	29.1	31.0	30.0	28.4	39.6

Table 2: Model performances on 500 samples when evaluated against the union of the "best" annotations ($B1 \cup B2$) and intersection of the "valid" annotations ($V1 \cap V2$), with the original question excluded. The difference between all numbers except the random and bag-of-ngrams are statistically insignificant.

3 Experimental details

Preprocessing: We tokenize the raw text in our post, question and answer using the NLTK tokenizer. We restrict the post to its first 300 tokens and the question and answer to first 40 tokens.

Word embedding model: Each post, question and answer in our dataset is represented using embeddings. To generate these embeddings, we train 200 dimensional word embeddings using GloVe on the 3 billion token datadump of StackExchange. We use a threshold frequency of 100 to create our vocabulary of 250,000 tokens. All tokens with a frequency of less than 100 in our dataset get assigned an 'UNK' token.

Model hyperparameters: The hidden layers in all the neural models are of size 200. We use ReLU non-linearity as our activation function between the hidden layers. We use a batch size of 128. We train the models for upto 14 epochs and at test time we use the predictions of the epoch where the performance on the tune set is the best.

Community QA baseline: We use the implementation[2] provided by the winning team of the SemEval2017 Community Question-Answering (cQA) subtask 3. Their original model contains six feature groups: string similarity features, word embedding features, topic modeling features, keyword features, meta data features and dialogue identification features. Since we do not have information about the latter three features in our dataset, we use only the first three features and train a logistic regression model to obtain the confidence scores on the positive labels.

Table 2 contains the results of the models when the original question is excluded (Section 5.2.3 in the main paper).

[2]https://github.com/TitasNandi/cQARank

4 Expert annotation details

We use Upwork for collecting our expert human judgments. Upwork is a platform which allows us to post a job description and recruit people specifically for a task. As a training process, we first ask the annotators to annotate a sample of 5 examples and provide them with feedback and additional guidance. We also ask annotators to rate their confidence in {1: Educated guess, 2: Pretty sure, 3: Quite sure}. The confidence on 17% of the annotations was rated as low, 47% was rated as medium and 37% was rated as high.

5 Example outputs

To understand the behavior of our EVPI model, we have include three example outputs in Table 3 one each from the three domains in our dataset. The first example is a case where the EVPI model predicts both the "best" and the "valid" questions higher in its ranking. The original poster is facing some issue they call the "suspend resume" issue. The post is unclear on what problem the poster is facing. Hence the "best" question asks for that information. In the second example, the model predicts one of the "valid" questions higher up in its ranking but fails to predict the "best" question. The model predicts "why would you need this" with very high probability likely because it is a very generic question, unlike the question marked as "best" by the annotator which is too specific. In the third example, the model again predicts a very generic question which is also marked as "valid" by the annotator. These examples suggest that the model is good at correctly predicting generic questions, but not at predicting very specific questions.

Title:	**Ubuntu 15.10 instant resume from suspend**
Post:	I have an ASUS desktop PC that I decided to install Ubuntu onto. I have used Linux before, specifically for 3 years in High School. I have never encountered suspend resume issues on Linux before until now. It appears that my PC is instantly resuming from suspend on Ubuntu 15.10 I am not sure what is causing this, but my hardware is as follows: Intel Core i5 4460 @ 3.2 GHz 2 TB Toshiba 7200 RPM disk 8 GB DDR3 RAM Corsair CX 500 Power Supply AMD Radeon R9 270X Graphics - 4 Gigs ASUS Motherboard for OEM builds VIA technologies USB 3.0 Hub Realtek Network Adapter Any help is greatly appreciated. I haven't worked with Linux in over a year, and I am trying to get back into it, as I plan to pursue a career in Comp Science (specifically through internships and trade school) and this is a problem, as I don't want to drive the power bill up. (Even though I don't pay it, my parents do.)
✓0.87	does suspend - resume work as expected ?
✓0.71	what , specifically , is the problem you want help with ?
✓0.70	the suspend problem exits only if a virtual machines is running ?
0.67	is the pasted workaround still working for you ?
0.57	just wondering if you got a solution for this ?
0.50	we *could* try a workaround , with a keyboard shortcut . would that interest you ?
0.49	did you restart the systemd daemon after the changes 'sudo restart systemd-logind' ?
0.49	does running 'sudo modprobe -r psmouse ; sleep 1 ; sudo modprobe psmouse' enable the touchpad ?
0.49	2 to 5 minutes ?
0.49	does it work from the menu or not ?
Title:	**Frozen Linux Recovery Without SysReq**
Post:	RHEL system has run out of memory and is now frozen. The SysReq commands are not working, so I am not even sure that /proc/sys/kernel/sysrq is set to 1. Is there any other "safe" way I can reboot w/out power cycling?
0.91	why would you need this ?
✓0.77	maybe you need to use your 'fn' key when pressing print screen ?
0.59	do you have sudo rights on this computer ?
0.55	are you sure sysrq is enabled on your machine ?
0.52	did you look carefully at the logs when you rebooted after it hung ?
0.51	i assume you have data open which needs to be saved ?
✓0.50	define " frozen " . did it panic ? or did something else happen ?
✓0.50	maybe you need to use your 'fn' key when pressing print screen ?
0.50	tried ctrl + alt + f2 ?
0.49	does the script process 1 iteration successfully ?
0.49	laptop or desktop ?
Title:	**How to flash a USB drive?.**
Post:	I have a 8 GB Sandisk USB drive. Recently it became write somehow. So I searched in Google and I tried to remove the write protection through almost all the methods I found. Unfortunately nothing worked. So I decided to try some other ways. Some said that flashing the USB drive will solve the problem. But I don't know how. So how can it be done ?
✓1.01	what file system was the drive using ?
1.00	was it 16gb before or it has been 16mb from the first day you used it ?
✓0.74	which os are you using ? which file system is used by your pen drive ?
0.64	what operation system you use ?
0.51	can you narrow 'a hp usb down ' ?
0.50	could the device be simply broken ?
0.50	does it work properly on any other pc ?
0.50	usb is an interface , not a storage device . was it a flash drive or a portable disk ?
0.49	does usb flash drive tester have anything useful to say about the drive ?
✓0.49	your drive became writeable ? or read-only ?

Table 3: Three examples of human annotations one each from the three domains in our dataset. The questions are sorted by expected utility, given in the first column. The "best" annotation is marked with black ticks ✓and the "valid"' annotations are marked with grey ticks ✓.