Questions for “Summarizing Answers in Non-Factoid Community Question-Answering” (Song et al., 2017):

1. Although there is a notion of relevancy when summarizing answers for a particular question, do you think that the authors are justified in not including a notion of correctness?
2. From the paper, it appears that the authors use n-grams from the answers directly during document expansion. Given that these are community answers, what effect will misspellings or bad grammar have on the results obtained?
3. If the quality of the community answers are poor, is there a chance that the features from the auxiliary dataset (Wikipedia) will dominate the generated summary?
4. For experimentation and evaluation, the authors use a hand-curated dataset with just non-factoid questions. Their methodology is tailored to be suitable for non-factoid questions. How well do you think the method will work for factoid questions?

- By Shreyas Udupa

1. **The author used coordinate descent to get the optimal value of the target function, does it make sense to use coordinate descent here?**
2. **The paper used CNN to generate additional text, but wouldn’t using some generative model more suitable for this task?**
3. **The author aimed specifically on non-factoid questions, while the performance on factoid questions remain not answered, will there model perform well on the later? (Moyuan Huang)**

1. **When representing sentence as a vector, why doesn’t the users use existed sentence embedding methods (e.g. sentence to vector, mean of pre-trained word embedding)?**
2. **In MMR algorithm, large κ results in high diversity while small κ results in high similarity. How do the author choose κ?**
3. **The author mentions that they ignore the semantic dependencies among the answers. Is there any study considering this aspect?**

(Zeng Fan)

1. **Can we use divrank instead of lexrank to improve the diversity in the summary?**

-Sudhanshu Bahety

1. The authors do not explain the reasoning behind choosing Wikipedia as the source for expanding the document? Will such choice lead to bias in modelling? Should the document expansion source be more diverse and board?
2. The dataset consists of only 100 non-factoid questions, how should the author justify the questions? How do these 100 questions be representative enough for a more broad spectrum of questions?
3. Many CQAs have answer rankings, does the results from this paper also reflect such rankings as well? Since while looking at the answers from CQAs, people tend to view high-ranked answers first.

By Yan Cheng

Questions for “Exploring Latent Semantic Factors to Find Useful Product Reviews” (Mukherjee et al., 2017):

1. The authors apply Latent Dirichlet Allocation (LDA) to learn the latent facets associated to an item. Online reviews are often short, and LDA is not known to work effectively on short documents and small global vocabularies. Is there a particular reason they chose this algorithm over one that could work better for variable length documents?
2. I feel they make many assumptions and implement a very complex model. Why can’t they make model comparisons to rank components of their algorithm by importance.
3. The related work and baselines are quite old. What is the current/future work for this problem?
Chester Holtz

1. They make an assumption that "expert users agree on what are the important facets of a product, and their description (or, writing style) of those facets influences the helpfulness of a review." Do you agree with this? There might be cases where experts have difference in opinion about an aspect’s importance esp. in cases of books/movies etc.
2. This is a 2017 paper and all the baseline models are from 2006-10. In those baselines too they compare after removing other features which could be the very basis of their models. Do you think that the paper could have selected better models for comparisons? Do you have any information on the performance of model compared to current state of the art.
3. Is the model biased towards frequent/expert reviewers as compared to good new reviewers? How can we try to correct this effect.

-Tushar Bansal

1. The paper proposed a general model across all domains of review. But from my point of view, you can always achieve better result on a narrow problem than on a general problem. Can their model be further improved by incorporating other knowledge, for example, the class specific knowledge?
2. Markov model is time consuming, how do you like the time-performance trade-off here?
3. In the quality assessment, the paper showed the stable increase of log-likelihood of their model. Do you think it is enough to valid their model's performance without showing the log-likelihood curve of other models?
   (Moyuan Huang)

1. The case of multiple users using the same account, doesn't seem to be considered by this model? Could this be added to the model, without bringing in too much complexity?
2. Majority of the reviews are written only when the consumers are dissatisfied with a product – and they write a critical analysis about varied facets of the item – and not just the ones advertised by the seller Moreover spam reviews would highly recommend the product by detailing the advertised features (this could be more than just opinionated) and could wrongly be interpreted as an expert review. Could this be accounted for especially for negative reviews?
3. This paper appears to consider user expertise as a time evolving factor and evaluates based on facets related to them, without giving much importance to domain related expertise. Could this be improved?
4. While the author considers items with >5 votes for the food category and for this case – the top words for amateur and expert reviewers don't seem to convey a lot. And although this model appears to perform better than the baselines it has been compared with, the model still appears to be lacking. So, in this regard, is the author's claim of generalization still valid?

(Balasubramaniam Srinivasan)

Questions for 'User Review Sites as a Resource for Large-Scale Sociolinguistic Studies' (Hovy et al., 2015)

1. How can we integrate this information/analysis of socio-economic impacts on language/linguistics to traditional topic-modelling approaches.
2. In many cases, we might not get this meta information from the data. Is it possible to learn these sociolinguistic features from one dataset and apply the learnings to other datasets. If yes, then would you have any ideas about how can we do that?
3. Authors claim that they plan to ‘relate the data to extra-linguistic information’ and ‘augment it with information on the grammatical information’. What do think exactly this extension intended for? And how can they do this?

- Tushar Bansal

1. In the beginning they talk about how the traditional datasets lack any sort of socio-economic data and they the web will help them get that. How does their data solve this problem ?
2. Can this data be used to model what kind of businesses people would be interested in depending on various factors like location, language, age, gender ?
3. A lot of users online tend to give fake names, age and other data. How do they filter that out ?

- Ajitesh Gupta

Extracting and Ranking Travel Tips from User-Generated Reviews (Guy et al., 2017)

1. If there is no gold set of domain-specific tips, how do we modify this mode?
2. The authors use people to evaluate, do you think we can use clustering to do this problem as semi-supervised problem?
3. It use the model to measure similarity with linear combining semantic similarity and word order, what if there is the interaction between them, could we add interaction term to measure the similarity?

   (Kuang Hsuan Lee)

1. A user’s opinion on what is considered a useful tip depends on his preferences and this model is purely on language features and a golden template decided by editors without any personalization whatsoever. How can we include some personalization into this model(Sai Chaitanya Kolasani)
2. Usage of english language may differ across geographical locations. For example the differences between british and american english are not trivial. Even in america there can be a difference in usage among different states. Is the

n-gram model with wildcard characters suitable for such differences ? (Sai Chaitanya Kolasani)

3. The authors mention that tips extracted from longer reviews tend to be more useful ? What is the intuitive explanation behind this ? Generally shorter reviews should be good candidate to extract tips as they are most likely on point with a clear intent. (Sai Chaitanya Kolasani)
4. The paper models tip as a small sentence in a review and authors put high emphasis on sentences starting with a infinitive verb. But generally tips are mentioned at the end of a long sentence. For example “We went to watch a movie here on saturday when It was very crowded and we didn’t have a pleasant experience, So avoid peak times.” Here avoid peak times is the tip but it is sort of at the end of a long sentence conveying user’s personal experience and also doesn’t start with a verb. So a better way would be summarization of entire text rather than hunting for short useful tips(Sai Chaitanya Kolasani)

1. This paper seems to take a more manual approach to analyzing the data and making predictions. Could we also apply either nonparametric language models (e.g. Naive Bayes) or a more expressive model such as LSTMs that is able to automatically model language and give useful results?
2. When performing tip ranking, could alternative choices of the similarity function between two tips lead to significantly different results than described in the paper?
3. How could we combine the tip candidate generation, classification, and ranking methods with other sources of data? For example, knowing something about the air quality in a particular location could influence how useful or actionable a tip is about that same location.

(Rajiv Pasricha)

1. Once the authors obtain the golden set of tips, why don’t they use to train a supervised method to map from review text to tips, if not for unsupervised methods, in order to generate more candidate tips ?
2. Is it efficient to use word2vec average of all words in the tip as the representation while measuring usefulness ? especially when there are tips containing over 15 words. The more words are added, the more the average vector points to the mean of the word2vec vector space. Wouldn’t it be better to use sequence based models or atleast top-k keywords for word2vec representation ?
3. Since each sentence is evaluated by a single editor, wouldn’t there be an inherent bias in the evaluation ?

- Ajitesh Gupta

1. If a review has sarcastic comment, how can you change the model to identify them?

- Sudhanshu Bahety

Neural Rating Regression with Abstractive Tips Generation for Recommendation (Li et al., 2017)

1. Do you think why the authors used tanh to do non-linear transformation instead of LeRU? Does the tanh works this situation better?
2. It transfers rating value to a rating vector, do you think what is the benefit to do?
3. The authors said W has too many parameters, maybe the authors can use W = Y*Y.T to reduce the number of parameters, what do you think the difference between this method and the neural network it provides?

   (Kuang Hsuan Lee)

1. How much of an effect does the review generation learning have on the solution here? Since they are using a simple neural network, wouldn't the tip generation architecture alone be able to model the required nonlinearities, given that the user sentiments in the review and tips would be quite similar, only more concise ?
2. If it is indeed useful, would it be better to use a proper generative model for the review generation process like an RNN/LSTM/GRU, instead of training it like a classifier as is done currently ? Especially since there would be thousands of words to be classified from.
3. While testing what was the criteria behind choosing 10 latent factors for NMF, PMF, SVD++ etc, while choosing 300 latent factors for proposed technique?

- Ajitesh Gupta

1. They use different embeddings for user and items but later project them into same hidden space. Instead of projecting user latent vector and item latent vector to same hidden space, it might be better to capture interaction between user and item at early stage. (Factorization machine / concatenate user and item latent vector)
2. Since they are not using reviews during the test time, they have a very poor context to generate tips. This model can generate very generic tips for an item which mostly depends on rating. These tips might not be useful nor actionable.

- Kulshreshth Dhiman

Detecting Evolution of Concepts based on Cause-Effect Relationships in Online Reviews (Zhang et al., 2016)

1. In section 5.3.1 - Grouping by similar concepts, they sort the pairs of clauses based on similarities and start from the most similar pair. For example, say (x1, x2), (x1, x3), (x5, x8), … are the pairs. Now, it could be possible that the x2 and x3 are not similar to each other, but since they are both similar to x1 in some way, they will be put into the same cluster according to the algorithm. I did not understand this part. In such a case where items are similar to one item in the cluster, but not the others, why are they assigned the same cluster?
2. The authors say that their methods can be applied to other scenarios as well. Would it not depend on the context in that case? For example, if you are studying a bigger dataset that consists of a whole variety of things, and you were studying the evolution of the term 'Ice Cream Sandwich', the context would have changed when the Android mobile OS having the same name was released, resulting in an abnormal rise in the frequency, and also a change in the cause->effect relationship.
   Aditi Ashutosh Mavalankar

1. How does the model tackle the issue that there weren't as many people using the internet in the early timeline as there were towards the end of it, causing a lack of information early on ?
2. Can the fluctuations in the word representation not just be caused by the fact that new words being learned were being learned over the course of time ?
3. How would you propose to form a set of cause->effect relationships in order to extend this approach to a broader set of technology ?
4. Could the results of this research be used to analyze why certain products failed/succeeded ?

- Ajitesh Gupta

Questions for '**Ask the GRU: Multi-task Learning for Deep Text Recommendations (Bansal et al., 2016)**'

1. What is the intuition behind using bidirectional RNN ? For english sentences it makes sense that words in the future may depend on the context from previous words. But how exactly a reverse RNN helps ? (Sai Chaitanya Kolasani)
2. The tags from the CiteUlike dataset are generated by users for their personal use and is inherently very noisy. How does the secondary learning task deal with this noise. (Sai Chaitanya Kolasani)

3. The proposed model alleviates the cold start problem for new items but it doesn't work for new users. The authors mention adding textual features of users to this model is very easy. What kind of textual features the author is referring to ? Generally websites don't have textual descriptions of individual users. (Sai Chaitanya Kolasani)

1. The model makes the assumption that each word in a document's abstract has equal importance in the prediction task. Since they also mention that they dont remove stop words, wouldn't an attention mechanism let the words speak for themselves and improve the performance? (Vignesh Gokul)

1. How is the cold start performance consistently better than the warm-start performance ? Shouldn't it be the reverse case ?
2. How does the author propose to use this model in case of non-sequential data such as images ?
3. As far as i understood, the authors rely on the assumption that latent factors learned for items and users through collaborative filtering is accurate, as in the learning process they are tuning their learned representations to fit with these factors. Isn't that an inaccurate base to start from ?

- Ajitesh Gupta

Wednesday Papers

Characterizing and Predicting Enterprise Email Reply Behavior (Yang et al., 2017)

1 - The model proposed in the paper has a lot of manual feature engineering. Can we instead propose a model with latent features that could automatically extract the best features that explain most variance in the data? This might work better as well.

2 - Do you think that the baselines considered in the paper are the best ones given that they seem very naive and are just based on heuristic? For example, authors could have used a related work they cited which just models the dyadic relationship as the baseline.

3 - **Previous reply** baseline generates predictions according to the previous reply behaviors of the recipients {u_j} towards the sender u_i before the sent time t of email m. However, since authors just aim to predict the reply towards first messages, they discard all the following replies in the thread (Page 3, last para). Wouldn't **Previous reply** be at disadvantage because of this, as subsequent replies could have provided further evidence to boost its performance?

-- by Rishabh Misra

1. The paper describes feature engineering related to various characteristics related to email. It relies on factors like whether email is tagged important/high priority. Will it be good idea to rather use Natural language understanding on text of e-mail to detect a sense of urgency? This can potentially help us avoid cases where a non-urgent mail is incorrectly marked as high priority by a sender.

2. Will size of attachment be useful feature? As many senders include a small image in their signature which might get counted as attachment. Large sized attachments can signal a paper/presentation which might be more relevant.

3. How should we handle the cases when the receiver is a broadcast list instead of a user?

By Nitin Kalra

1. Can we approach the problem in a generative way? I think the paper proposed many hand-curated features, can we just throw all the features to a CNN for example and let the model to figure that out?
2. During reading the paper, I feel like many features are very straight forward, like the time of the day email is sent. For me, I would already know that a email sent around 2~4 are likely to be replied since people would check their email before work is over. So, how do you like the contribution of this paper?
3. I think the baseline models selected is too naive and simple, is there a better model in this field?
(Moyuan Huang)

1. Do you think if we put all text content into neural network and transfer these words to some hidden features, does it get better result?
2. The authors focus on the feature engineering and these features based on the data analysis on the specific data, it kinds of not generative? How do you think it?

3. The number of sent emails in each month is not uniform distribution, do you think maybe we should select the more smooth data instead of the data affected by layoff event?

   (Kuang Hsuan Lee)

1. The training dataset are emails in almost twenty years ago. Will this affect the method of performance?
2. Enterprises usually have calendar systems. Predicting reply time without considering the calendar is non-realistic. Is there any way to incorporate the calendar information into the method?
3. Will the difference in the number of training samples with replies and without replies affect the method's performance?

   (-Siyu Jiang)

1. The proposed model has been applied to one of the tech companies. Can we generalize this model to work for different industries say fashion, retail, etc?
2. Can the sender's level within the organization be used as a feature? For example, an employee is more likely to reply to an email sent by his manager than his peer.
3. Do you think this model should also consider spam filters which will redirect email to the spam folder than the inbox?

   (Prem Nagarajan)

1. How well do we expect the approaches used to predict reply behavior to generalize beyond the Avocado dataset?
2. Would the performance of the algorithms improve if the authors employed some kind of mechanism to detect highly correlated features? It seems that many of the features are likely to be highly correlated which may interfere with the performance of some algorithms.

3. What techniques could be used to automatically construct useful features from emails? Would it be useful to automatically learn more descriptive text features rather than just a bag-of-words representation of the email body? Something like PCA could help create more applicable features for the specific prediction tasks.

Rajiv Pasricha

1. The data source used in this paper consists of only 279 accounts while, in real life, enterprises tend to be huge and have hundred of thousand of employees. How do the authors justify and validate the correctness and representativeness of the result in real life situation?
2. How do the authors take the external influences of emails into consideration? For examples, considering the majority of enterprises, there are sure to be influences from product launching. When those days are approaching, there is sure to be a rise in email amount and rise in reply rate. How do this paper adjust such conditions into their model?
3. How do the authors fulfill the future research that will consider more available email collections given the reality that enterprise emails are generally regarded as business secret and shall not be shared in public?

By Yan Cheng

Understanding How People Use Natural Language to Ask for Recommendations (Kang et al., 2017)

1. Does this approach apply to music, or shopping domain?
2. Do you think the results can apply to other country, maybe culture different?
   (Kuang Hsuan Lee)

1. Can we capture emotions of the users through voice better than the text based systems? Therefore providing better recommendations.
2. Are there more follow-up queries for people who are non native english speakers?
3. Are there any scenarios where speech based queries perform significantly better or worse than text?

(Prem Nagarajan)

1. The author used a movie site for data, will this cause bias or faulty conclusion? Since it is very common that people tend have different standard when talking about different things. For example, people may be really willing to listen to different kind of music but they may stick to one kind of movies. Will such different influence the model here?
2. The sample site seems relatively small, is this going to lead to bias? Also, the model does include social-economic metadata in consideration, will different gender, age, or social-economic class cause bias in the study?
3.The study mentions that many subjects reformulate their query or start over, but does not mention the accuracy of wit.ai's translation. Will such reformulations of voice search result in the inaccurate translation from wit.ai?
By Yan Cheng

1. How important is the demographic of the participating user to this study? - gender, age, primary language etc. (Kiran Kannar)
2. The authors state that they collected one response per subject. Is this not too small to understand interactions and the use of natural language? (Kiran Kannar)
3. An important aspect of interaction studies is the testing of UI design. While the authors mention that they improved the prompt wording, could they have done any better or differently? The Google-style interface may not have been familiar in this integrated set-up, while it is extremely familiar with Google UI. (Kiran Kannar)
4. How can we model the intent of an user? The paper is more of a observational paper, than a modelling one. (Kiran Kannar)

1. The number of sample they used is pretty small. Although they did some filtering and selecting of users, could this dataset show our truth?
2. If not, how can we collect a larger and better dataset to do this behavior analysis and get deeper understanding of the usage of natural language to ask for recommendation
3. I think the speech recognition they used is not perfect, so some mistakes in this process will introduce biases to the interaction with user and then to the feedbacks from user. Are they detect those things?
(Wen Liang)

1. How useful is it to treat refinement and reformulation queries differently, given that the second query is mostly an addition to the first one except for the fact that in reformulation queries the original question is repeated ?
2. Even the examples of refine with constraints and refine with clarification do not clearly demarcate the differences between them. Like is there much difference in saying “more sitcom less absurd” vs “More true horror instead of drama/ thriller” ? What would be more concrete rules to identify the required differences ?
3. The authors address the fact that they have used a very small sample of users in their data and among them some are using voice while others use text, and as such it may have biases. But aren’t they introducing bias by themselves by keeping such vague definitions of query types and as leaving them open to interpretation by the 2 people who classified them ?

- Ajitesh Gupta

Questions for “Smart Reply: Automated Response Suggestion for email” (Kannan et al., 2016):

1. For the triggering part, the model does not consider if other people are involved in the email thread or not, and suggests responses regardless of who the message is addressed to. How and at what stage do you think such a notion can be incorporated into the model?
2. “First, to improve specificity of responses, we apply some light normalization that penalizes responses which are applicable to a broad range of incoming messages”. What form of 'normalization' could the authors be referring to?
3. To include negative responses with the positive responses, a second pass over the LSTM network is performed. Why can't a few lowest scoring responses be saved from the first pass?

- By Shreyas Udupa

4. **During semi-supervised learning, there is a constructed graph with feature nodes. Is it possible to convert the feature nodes to embedding to learn the semantic labels?**
5. **In the trigger network, why do the authors choose to use unigram/bigram feature instead of word embedding?**

6. How does this model normalize the responses?

(Zeng Fan)

1. During the implementation, What kind of normalization is the author using to penalize for the broad messaging?
2. How much does the normalization effect the final outcome? For the broad range of message, does it make big differences if we are just ‘slightly’ normalize it?
3. I don’t quite get the 2 passes LSTM, I think the responses should be able to be handled in a single pass, considering the time cost for training it. Do you think there is any way to improve this?
   (Moyuan Huang)

1. This work clusters responses of similar meaning, but is there a way to deal with different tones? For example, “Yes, I will” and “Yes, I am glad that I can help” are both affirmative responses, but they convey different level of politeness.
2. How will this work extend to more neutral responses? In this work, most responses are either affirmative or negative. But in many real emails, the responses are not limited to yes and no.
3. How to decide what kind of seed semantic intent are needed during training?
   (-Siyu Jiang)

1. The paper presents idea of manually defined clusters with seed examples. How should the size of clusters/seed examples be determined in such an approach?

2. The paper talks about enforcing negatives and positives. How should we deal with neutral responses? There can be questions, say about politics, etc. for which neutral responses might be most popular.

3. Even though the triggering module might filter a message as bad candidate, can we somehow use such a message in order to expand the variety of good candidates in future, maybe with the help of actual response of user?

By Nitin Kalra

1. The smart reply provided by Google is short and very simple. However, compared to messages and chat, emails are usually used when we are trying to write a longer piece of message and give much details about what we are trying to communicate. Also, emails are generally tend to be more formal. Will such smart reply blurs the distinction between message and emails?
2. When the author mention that they normalize the incoming message and penalize certain ones. How do they normalize it?
3. Given people’s habit of writing being different, smart reply is lacking such personalities. People all have different writing styles and habits, and smart reply is

erasing such differences among different people. Is there any way for the smart reply being more custom and personalized?

By Yan Cheng

1. Can this smart reply feature be incorporated for instant messaging applications like WhatsApp where the reply needs to be generated instantaneously?
2. Most of the time the smart reply generated by gmail is very small which is orthogonal to the way emails work. Do you think the smart reply feature by itself cannot send a complete email and user editing is needed invariably.
3. Can this smart reply feature work for business emails, say internal emails of a technology company like Google or is it suited mostly for personal emails?
   (Prem Nagarajan)

1. For the response about date or time, could we improve the model by just giving a structure of response and let user fill in some time?
2. Do you think they can broadcast the suggestion diversity part to provide different levels of formality of email writing.
3. This model can only provide very short responses and exclude some situations, could we use similar way to generate a longer response with more specific information, more complex or just generate a basic structure of a response?
   (Wen Liang)

1. How do you think this model can be tweaked to account for user-specific language?
2. Is there a system that tests for the correctness of the replies, not in terms of syntax, but also semantics? And if the system has access to the user's personal data, such as contacts, calendar, etc., would that not be a privacy breach?

3. How do you think this system can be extended with respect to the language detection stage, where the non-English messages are discarded? Does it mean that messages having a single non-English word are discarded? Would that also discard names that are not common in the English language?

- Aditi Ashutosh Mavalankar

Novelty based Ranking of Human Answers for Community Questions (Omari et al., 2016)

1. **The author used manually annotate the dataset, do you think could we use semi-supervised method?**
2. **Did the author choose the questions randomly or manually?**

   (Kuang Hsuan Lee)

1. **When doing proposition filtering, the author used CQA-ESA, which contains past CQA questions and their best answers. The author claims that this is to represent the similarity of texts which are different in languages. What if the past CQA collections do not contain relevant information of the query questions and answers?**
2. **The dataset used in this work contains only 110 questions and 1426 answers. Is this a rather small dataset for the CQA task? The average number of answers for each question is only 13, and each answer is not long (as mentioned in the paper). Most users can easily read most of the answers. Why is ranking necessary in this scenario? Would this dataset fully reflect the performance of the methods?**
3. **I think in CQA, novelty is a non-primary ranking metric. Currently, common CQA platforms such as Quora or Zhihu usually use the number of votes for an answer from users as a primary ranking metric. Is there any method to combine the votes with this work's novelty based ranking?**

(- Siyu Jiang)

1. How do you think their method would perform on short question/answer pairs vs. longer pairs present in their dataset?
2. Do you think it is more valuable to learn under what contexts novelty is prefered over answers with fewer novel propositions?
3. Can you discuss what advantage did the manually annotated data gives?
Chester Holtz

1. While the idea behind the paper is great, I feel that the stumbling block is the issue of defining the similarity between questions and answers. Computing the similarity between short pieces of text without additional context is still an unresolved problem. Is there any way the authors can bypass the computation of similarities in this task?
2. The authors do not address the recency of answers. In Q/A forums such as StackOverflow, a more recent answer might usually contain more up-to-date information. How can this model be extended to account for recency?
3. Since the authors assume that the "wisdom of the crowd" is a good thing, can the votes earned by each answer be used to train a learning to rank model?

Siddharth Dinesh

1. If the CQA are very short, then it would be difficult to capture the semantic aspect. How can you modify the approach to combat this problem?

-Sudhanshu Bahety

1. In proposition filtering, we use 2 million CQA from yahoo as documents for the latent space. Is there any efficient way to represent them in lower dimension?

- Digvijay Karamchandani

1. The paper randomly selects 2 million CQA yahoo answers for representing the latent space. Is there any better way to select documents maybe clustering using K-means and then using K latent space?

-Kriti Aggarwal

1. Could we combine this model with some other features such as user history, current upvote, upvote rate, comments?
2. Could we rank the answers personalized to different group of users? For example, a novice want some answers which are easy to understand others want deeper understanding.
3. Some answer or comments are outdated but still have some good insights. So Is it possible to make time-variant selection and summarization?
   (Wen Liang)

They collect CQA questions and their best answers from 2 million Yahoo Answers

- Did they filter out questions with few answers?
- Best answer may ensure relevancy to the question. How does this ensure diversity of answers? Why not take top-K answers (having significant #upvotes)?

--Kulshreshth Dhiman

Application: Efficient Natural Language Response Suggestion for Smart Reply (Henderson et al., 2017)

1. What challenges would the model need to tackle to incorporate multi-language email-based smart replies. (Kiran Kannar)
2. The conversion rates presented are relative to Seq2Seq. And the model only seems to perform slightly better than Seq2Seq. Are we only looking at computational advantage then? (Kiran Kannar)

1. Is it possible to extend this one step further by personalizing replies according to the general linguistic profile of the user ? if so what could be some possible ways to go about it ?
2. Can the diversification process be introduced with the process of response formulation itself ? Like during beam search or at training level to give importance to diverse responses ?
3. How do they tackle issues of specificity like in case of suggesting “Lets do it on Thursday”, how do they decide that thursday would be a good option ?
4. How do they tackle sensitive replies which can be answered in short ? Like for an extreme case when a user gets an invite for a sad event, how do they learn not to give overly positive suggestion such as “Yeah that would be great” especially when they say that most generated responses are positive ? is it handled by the RNN or by the triggering mechanism ?
5. Would it be beneficial to involve context of the whole email thread rather than treating each email individually ?

- Ajitesh Gupta