Question Answering over Electronic Devices: A New Benchmark Dataset and a Multi-Task Learning based QA Framework

Supplementary Material

1 Introduction

The supplementary is organized in the same sectional format as the main paper. The additional material of a section is put in the corresponding section of the supplementary so that it becomes easier for the reader to find the relevant information.

Some sections and subsections may not have supplementary so only their name is mentioned.

2 Corpus and Datasets

2.1 Creating the E-Manuals corpus used for pre-training

Pre-processing of Pre-training Corpus: Each PDF is read in a hierarchical manner (PDF → block → span) to keep the order of the text intact, and the images are ignored (if any). The 'PyMuPDF'[1] python package is used for reading the PDFs. We remove the table of contents and all the non-Unicode and non-ASCII characters from the E-manuals. We concatenate the cleaned text of all the E-Manuals, thus collecting a total of $11,653,755$ paragraphs, each having an average of 4.4 sentences.

Sample paragraph from Pre-training Corpus Two sample paragraphs from the corpus are as follows (these samples show that the text in the corpus is mostly instructional) -

`"1. While the printer is idle, press the Help pages menu item. 2. Note the IP address on the print and save the print for later reference. Leave the printer plugged into its power outlet; this preserves a ground path for static discharges. Touch the printer's bare metal frame often to discharge static electricity from your body. Handle the circuit board(s) by their edges only. Do not lay the board(s) on a metal surface. Make the least possible movements to avoid generating static electricity. Avoid wearing wool, nylon or polyester clothing; they generate static electricity."`

`"Batteries Warning Batteries should never be exposed to flame, heated, short-circuited or disassembled. Do not attempt to recharge alkaline, lithium or any other non-rechargeable batteries. Never use any battery with a torn or cracked outer cover. Keep batteries out of the reach of children. If you notice anything unusual when using this product such as abnormal noise, heat, smoke, or a burning odor: 1 remove the batteries immediately while being careful not to burn yourself, and; 2 call your dealer or local Olympus representative for service. AC Adapter"`

Word Cloud characterizing pre-train corpus

Fig. 1 shows a word cloud for the top 200 most frequently occurring words in the above two paragraph samples. Red boxes enclose verbs that bring out the instructional and assertive nature of the sentences.

[1] https://pypi.org/project/PyMuPDF/

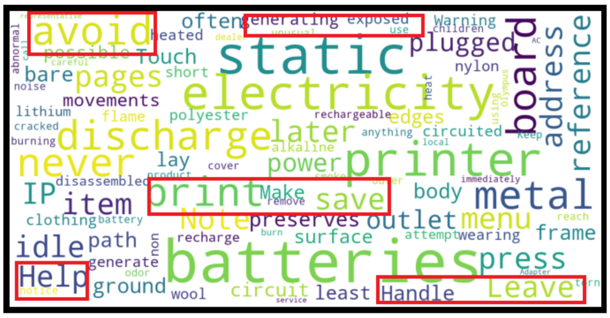

Figure 1: Word Cloud for the most frequently occurring terms in sampled paragraphs. The red boxes enclose verbs containing the instructional and assertive nature of the sentences. eg: "avoid", "help", "handle", "leave", "print".

2.2 Question Answering Dataset from E-Manual

Samsung brand was chosen for the following reasons - (1) Samsung manufactures a variety of models of smartphones, Televisions and other electronic goods. (2) E-Manuals of Samsung are easily available[2] in HTML as well as PDF formats, and are very well organized. (3) A large number of user forum questions are available in Amazon, which makes study of consumer question forums possible. However, other popular brands could also be chosen, we believe that would not make much of a difference. (4) According to Gartner, Samsung is ranked 1 in terms of Digital IQ [3] [4] [5], which may be treated as a proxy of how a brand is able to integrate in the smart technology ecosystem.

Analyzing quality of Annotated Question Answering Dataset

In order to evaluate the quality of the expert annotations, we use the crowdsourcing platform Appen [6] to launch two crowdsource surveys - one of S10 QA Dataset and other for Smart TV/Remote QA. 100 QA pairs each are randomly sampled from the S10 QA and the Smart TV/Remote QA Dataset separately, and corresponding to these pairs, questions, sections containing their answers, the answers annotated by the expert annotators and the E-Manual are given to crowdworkers. Each worker needs to decide if the annotated answer satisfactorily answers the corresponding question. 3 judgements are considered per question, and the workers that finish an annotation in less than 3 minutes are flagged, thus avoiding spam. The crowdworkers answer using an interface illustrated in Fig. 2.

Also, there are three levels of crowdworkers mentioned in Appen - **Level 1** - Fastest Throughput: All qualified contributors **Level 2** - Higher Quality: Smaller group of more experienced, higher accuracy contributors **Level 3** - Highest Quality: Smallest group of most experienced, highest accuracy contributors We select the 'Level 3' of crowdworkers to ensure that the annotation quality of the crowdworkers performing the survey is not compromised with. Table 1 shows the results of the survey, showing that for more than 95% of the samples, majority of crowdworkers agree with the expert annotation in both the surveys. Also, the quality of the surveys in terms of clarity and ease of job is quite good based on the ratings given by some crowdworkers.

[2] https://www.samsung.com/us/support/downloads/
[3] https://www.gartner.com/en/marketing/insights/daily-insights/top-10-consumer-electronics-brands-in-digital-3
[4] https://www.gartner.com/en/marketing/insights/daily-insights/top-10-consumer-electronics-brands-in%2Ddigital-4
[5] https://www.gartner.com/en/marketing/research/digital-iq-index-consumer-electronics-us-2020
[6] https://client.appen.com/

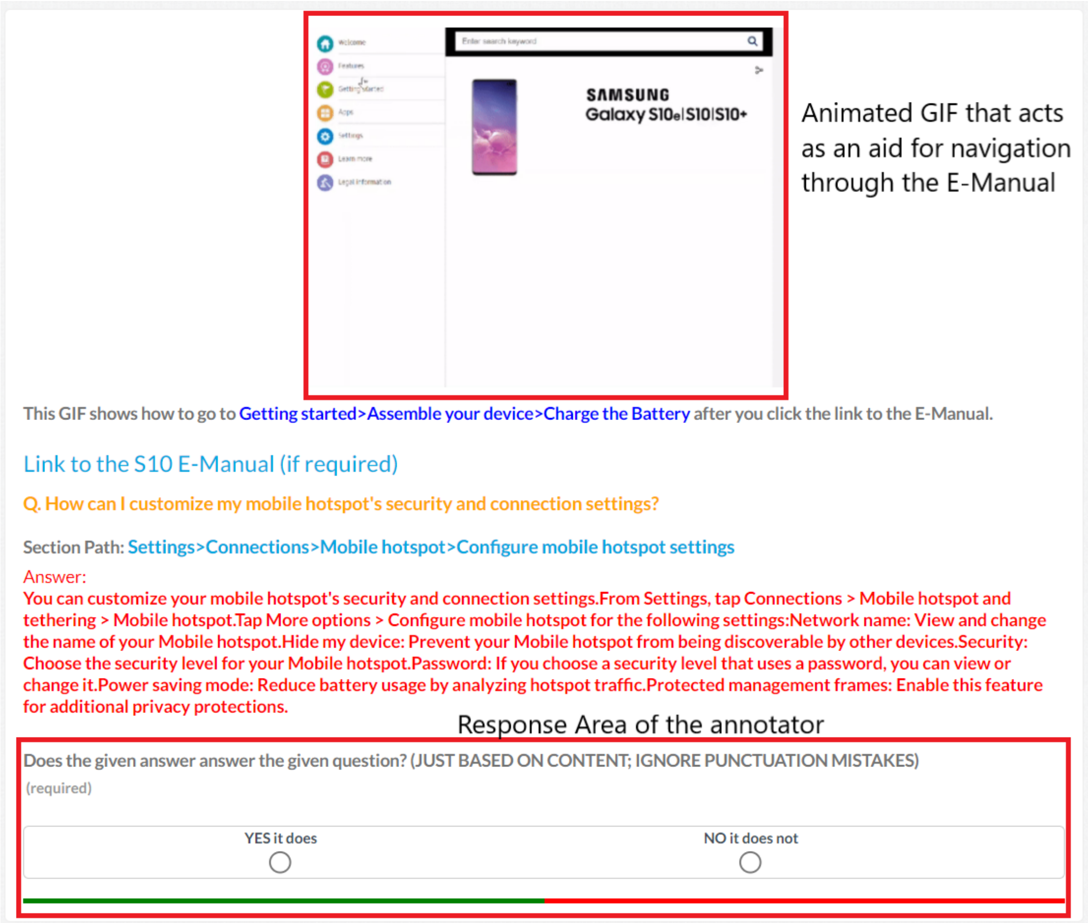

Figure 2: User Interface for the crowdworker

Measure of the agreement between crowdworkers and experts	S10	Smart TV/Remote
No. of crowdworkers (excluding flagged ones)	116	210
No. of randomly chosen samples from the S10 QA Dataset	100	100
No. of samples where all crowdworkers agree with each other and the expert	73	76
No. of samples where majority of crowdworkers agree with the expert	96	100
Quality of the crowdsource survey as rated by some crowdworkers	S10	Smart TV/Remote
No. of crowdworkers who rated	13	8
Average rating for clarity	3.6/5	4.5/5
Average rating for ease of job	3.3/5	4.3/5

Table 1: Results of the crowdsource survey

Comparison between TechQA and S10

The size of our datasets is comparable to that of the TechQA Dataset (which belongs to the Technical Support Domain and hardly contains questions pertaining to electronics consumer products). Our datasets have **small question lengths, long answer lengths and answers that have multiple spans**, which makes it different from TechQA dataset. Also, the distribution of the number of tokens per question in our datasets is similar to that of a set of 1028 Questions extracted from Amazon Question Answer Forum when comparing the range (approx. $5 - 15$) that comprises most of the density, as can be seen in Fig. 3, thus making our annotated datasets a suitable proxy for Consumer Question Answering Forums. However, a significant portion of the distribution of the question lengths in TechQA Dataset is spread over a larger range (hence truncated in Fig. 3), and is very different as compared to that of Amazon Question Answering Forum. If we consider the way that the domain-specific TechQA Dataset was curated, the questions were taken from technical forums, and answers from technical documents. However, we ask annotators to frame questions themselves from E-Manuals, by marking the answer first, and then framing the question. This would make the question set more answerable, and the questions thus obtained would be of better quality.

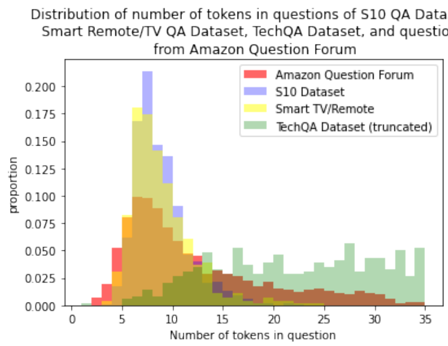

Figure 3: Comparison of normalized distributions of tokens per question of S10 QA Dataset and a set of questions extracted from Amazon Question Answering Forum

2.3 Questions from the real consumers

2.4 Questions spanning across several devices

Sample Questions for analysis on other devices

These are the 10 sample questions that were asked across several devices -

1. Does it use a sim card?
2. How do I switch off the device?
3. Does it use a SD port?
4. Does this device offer Wi-Fi calling ?
5. How can I change the device language ?
6. How can I set the brightness level ?
7. How can I hide the notifications ?
8. How can I change the Font size ?
9. How can I use stopwatch?
10. How do I setup tones on my device?

Question Paraphrase Detector: This is used for detecting Amazon User-Forum Questions that are answerable, by detecting whether it is a paraphrase of the most similar Annotated Question or not. For this, the CQ-AQ Paraphrase Dataset is split into train, validation and test sets in the ratio of $8 : 2 : 1$ for training and evaluating a **question paraphrase detector** - this is a RoBERTa Sequential Classification Model (initialized by weights of RoBERTa pre-trained on E-Manuals), as shown in Fig. 4. This method gives a high precision of 0.932, and a high recall of 0.814.

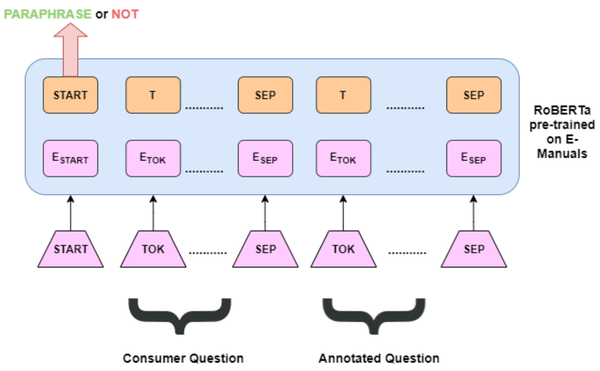

Figure 4: Question Paraphrase Detector

3 Methodology

Overview of Pipeline

The EMQAP is laid out in the form of a pseudo-code in Algorithm 1.

Retrieving top k sections

Given an E-Manual, our first step is to reduce the pipeline's search space and provide it with only a few candidate sections for a question. We use an unsupervised IR method that accepts a question and all sections of the E-Manual as input and provides similarity scores for each question-section as output. We select the K highest scoring sections, which possibly contain the answer. Experiments show that the best way of representing question-section is by TF-IDF. Thus we create TF-IDF vector representations of questions and sections and calculate the cosine similarity of each question-section pair.

However, we make an enhancement by augmenting a section with probable questions that can be answered by that section (Nogueira et al., 2019). These questions are generated by a pre-trained T5 (Text-to-Text Transfer Transformer) (Raffel et al., 2020) model, which takes the section as input and outputs a list of questions that are answerable by that section. This augmentation results in the re-weighting of the terms, especially the terms which act as anchor when questions are framed receive more weights. We find this leads to improved retrieval of top k sections. We name this improvisation as TF-IDF + T5.

3.1 Pre-training on the E-Manuals Corpus

State-of-the-art pre-training of transformer models include masked language model pre-training (Devlin et al., 2019; Liu et al., 2019), next sentence prediction (Devlin et al., 2019), elastic weight consolidation (EWC) (Kirkpatrick et al., 2017), a decaying learning rate as a function of layer depth (Arumae et al., 2020), using heuristic data selection methods for an experience replay buffer (de Masson d'Autume et al., 2019), etc. Also, domain-adaptive fine-tuning methods have been used for transformer language models pre-trained on generic data such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) in order to improve performance in downstream tasks such as sequence labelling (Han and Eisenstein, 2019), duplicate question detection (Rochette et al., 2019) etc. Ramponi and Plank (2020) suggests unsupervised domain adaptation methods, that do not even require domain-specific annotated data.

Justification behind using masked language modeling We did not use the Next Sentence Prediction (NSP) pre-training task (Devlin et al., 2019), as it has been shown in Liu et al. (2019); Yang et al. (2019); Joshi et al. (2020) that NSP worsens performance in downstream QNLI (Wang et al., 2018) tasks and in question answering on the SQuAD Dataset (Rajpurkar et al., 2016). Also, intuitively, sentences in E-Manuals sometimes do not have dependencies with an adjacent sentence. Instead, there might be many sentences that are dependent on a particular statement that is not necessarily adjacent, as shown in Fig. 5.

Justification behind having a single epoch iteration. We pre-train RoBERTa on E-Manuals only for 1 epoch. This is as per the justifications put forward by Komatsuzaki (2019). (1) Single epoch ensures

Algorithm 1: EMQAP Pipeline

```
Function Pre-Training(corpus, RoBERTa):
    model = initializeWeights(RoBERTa, weights from (Liu et al., 2019))
    pre-trainedModel = MaskedLanguageModeling(model, corpus)
    return pre-trainedModel

Function MultiTaskLearning(pre-trained-model, Annotated-QnA, E-Manual)):
    copy-weights(pre-trained-model.encoder, supervised-IR.encoder)
    copy-weights(pre-trained-model.encoder, supervised-RC.encoder)
    //batch fine-tuning
    for QnA-batch in Annotated-QnA do
        questions, annotated-answers = QnA-batch
        topK-sections-batch = [unsupervised-IR(question, E-Manual) for question in
          questions]
        IR-prediction = supervised-IR(questions, topK-sections-batch)
        RC-prediction = supervised-RC(questions, topK-sections-batch)
        IR-Loss = Loss-Function(IR-prediction,
          sections-containing-annotated-answers)
        RC-Loss = Loss-Function(RC-prediction, annotated-answers)
        Loss = average(IR-Loss, RC-Loss)
        Back-propagate(Loss, supervised-IR, supervised-RC)
    end
    return supervised-IR, supervised-RC

Function Main():
    extract listOfEManualURLS from www.manualsonline.com
    corpus = createCorpus(listOfEManualURLS)
    pre-trainedModel = preTraining(corpus, RoBERTa)
    supervised-IR, supervised-RC = MultiTaskLearning(pre-trainedModel,
      AnnotatedQnA, E-Manual)
    //inference, given a question and the E-Manual from which the question is asked.
    topK-sections = unsupervised-IR(question, E-Manual)
    pred-section = argmax(supervised-IR(question, topK-sections))
    pred-answer = hard-classifier(supervised-RC(question, pred-section))
    return pred-answer
```

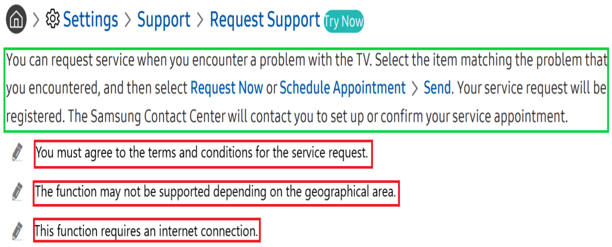

Figure 5: A sample from an E-Manual. Although the sentences enclosed by red boxes are adjacent, they are independent of each other. Instead, each such sentence is dependent on the sentences in the green box.

better diversity in the samples processed as compared to multi-epoch training thus preventing overfitting (2) Sampling from the training data matches the underlying data distribution in single epoch (3). RoBERTa has about $125M$ parameters. In our case, the number of batches is close to $80,000$, and the number of tokens (T) in the pre-training E-Manuals corpus is close to $1B$, making the ratio $T/P \approx 8$, which satisfies the optimal conditions for pre-training for one epoch as per Komatsuzaki (2019).

3.2 A Multi-Task Learning Approach for SR and AR

4 Experiments and Results

Evaluation of unsupervised IR methods

We evaluate the performance of our algorithm **TF-IDF + T5** (*detailed in suppl.*) with different baselines.

Baselines: We evaluate several baselines, such as - (a). Jaccard Similarity (**Jaccard Sim**) and Word Count Vector Similarity (**Count Vec Sim**) between the tokens of a question and the sections. (b). **Cosine similarity** between averaged pre-trained neural word embedding vectors such as **word2vec** (Mikolov et al., 2013), **GloVe** (Pennington et al., 2014), and **FastText** (Bojanowski et al., 2017) of the tokens of a question and the sections. (c). **Cosine similarity** between the sparse vectors generated using **TF-IDF** on tokens of a question and the sections. (d). **Cosine similarity** between pre-trained neural sentence vectors like **InferSent** (Conneau et al., 2017) of a question and the sections.

	Hits@1	Hits@5	Hits@10
InferSent	0.033	0.1	0.156
Jaccarrd Sim	0.222	0.422	0.467
Count Vec Sim	0.333	0.6	0.633
GloVe Sim	0.256	0.567	0.711
fasttext_sim	0.356	0.711	0.756
word2vec_sim	0.333	0.711	0.767
TF-IDF	0.511	0.889	0.911
TF-IDF + T5	**0.533**	**0.911**	**0.934**

Table 2: Unsupervised Information Retrieval Methods evaluated on S10 QA.

Results: We evaluate Hits@K that is, the fraction of the number of times the section relevant to a question appears in the top K sections for the baselines and (**TF-IDF+T5**) and report the results in the Table 2 for the test set of 90 questions of the S10 QA dataset. As can be seen **TF-IDF+T5**, gives the best Hits@K for $K = 1, 5, 10$ values = 0.533, 0.911, 0.934.

4.1 Evaluating MTL Framework

Table 3 shows three examples of questions and the corresponding predictions of EMQAP (Sentence-Wise Classification) and baselines.

Question	How can I turn on and turn off fast wireless charging?	Where can I find an option to setup separate app sound?	What is Samsung DeX for PC?
Ground Truth Answer	From Settings, tap Device care > Battery for options. Fast wireless charging - Enable or disable fast wireless charging when using a supported charger.	You can play media sound on a speaker or headphones separate from the rest of the sounds on your device. Connect to a Bluetooth device to make this option available in the Audio device menu. From Settings tap Sounds and vibration > Separate app sound. Tap Turn on now to enable Separate app sound and then set the following options - App > Choose an app to play its sound on a separate audio device. Audio device - Choose the audio device that you want the app's sound to be played on	Connect your device to a PC for an enhanced multitasking experience. Use your device and PC apps side by side. Share the keyboard mouse and screen between the two devices. Make phone calls or send texts while using DeX . samsung.com/us/explore/dex
EMQAP	From Settings tap Device care > Battery for options. Battery PowerShare - Enable wireless charging of supported devices with your device's battery. Fast cable charging - Enable or disable fast cable charging when connected to a supported charger	You can play media sound on a speaker or headphones separate from the rest of the sounds on your device. Connect to a Bluetooth device to make this option available in the Audio device menu. From Settings tap Sounds and vibration > Separate app sound. Tap Turn on now to enable Separate app sound and then set the following options - App > Choose an app to play its sound on a separate audio device. Audio device - Choose the audio device that you want the app's sound to be played on	Connect your device to a PC for an enhanced multitasking experience. Use your device and PC apps side by side. Share the keyboard mouse and screen between the two devices. Make phone calls or send texts while using DeX. Visit for more information - samsung.com/us/explore/dex
DPR	depending on device condition or surrounding environment	Settings	Volume. Tap More options > Media volume limit
MultiSpan	Enable	Audio device menu	enhanced, multitasking
TAP	Select a power mode to extend battery life. App power management : Configure battery usage for apps that are used infrequently. Wireless PowerShare : Enable wireless charging of supported devices with your devices battery. Fast cable charging : Enable or disable fast cable charging when connected to a supported charger. Fast wireless charging : Enable or disable fast wireless charging when using a supported charger.	make this option available in the Audio device menu. From Settings, tap Sounds and vibration > Separate app sound . Tap Turn on now to enable Separate app sound, and then set the following options: App : Choose an app to play its sound on a separate audio device. Audio device : Choose the audio device that you want the apps sound to be played on.	device to a PC for an enhanced, multitasking experience. Use your device and PC apps side-by-side Share the keyboard, mouse, and screen between the two devices Make phone calls or send texts while using DeX Visit samsung.com/us/explore/dex for more information.
Remarks	For complex procedural questions, EMQAP and TAP give the answer closest to the ground truth.	For 'where' type questions, (asking the location of a particular feature), EMQAP again performs very well as compared to the other baselines.	Factual ('what' type) questions are answered equally well by EMQAP and TAP.

Table 3: Examples of question-answer pairs from the Samsung S10 QA Dataset and predictions by EMQAP (sentence-wise classification) and baselines with remarks, explaining the predictions.

4.2 Evaluating Pretraining Techniques

We present three examples of different question types and their predictions and ground truths in Table 4 given by 2 variants and EMQAP, along with some remarks. We observe that EMQAP gives better answers for questions that inquire about procedure or location compared to variants. However, factual questions are answered similarly by all the models. Also, considering Table 5, we can see that EMQAP performs better than SQP(EWC+LRD) in all three categories, making a considerable improvement in answering location-based questions. Hence, we can say that questions regarding the device's operation and features are answered better by the EMQAP compared to all other variants. Also, the SQP(EWC+LRD) variant is better than the SQP(SLR) in answering the questions, which indicates the superiority of the training scheme. If we consider the questions containing non-contiguous ground truths, EMQAP performs better than SQP(EWC+LRD), as can be seen in Fig. 6.

Question	How can I turn on and turn off fast wireless charging?	Where can I find an option to setup separate app sound?	What is Samsung DeX for PC?
Ground Truth Answer	From Settings, tap Device care > Battery for options. Fast wireless charging - Enable or disable fast wireless charging when using a supported charger.	You can play media sound on a speaker or headphones separate from the rest of the sounds on your device. Connect to a Bluetooth device to make this option available in the Audio device menu. From Settings tap Sounds and vibration > Separate app sound. Tap Turn on now to enable Separate app sound and then set the following options - App > Choose an app to play its sound on a separate audio device. Audio device - Choose the audio device that you want the app's sound to be played on	Connect your device to a PC for an enhanced multitasking experience. Use your device and PC apps side by side. Share the keyboard mouse and screen between the two devices. Make phone calls or send texts while using DeX . samsung.com/us/explore/dex
EMQAP	From Settings tap Device care > Battery for options. Battery PowerShare - Enable wireless charging of supported devices with your device's battery. Fast cable charging - Enable or disable fast cable charging when connected to a supported charger	You can play media sound on a speaker or headphones separate from the rest of the sounds on your device. Connect to a Bluetooth device to make this option available in the Audio device menu. From Settings tap Sounds and vibration > Separate app sound. Tap Turn on now to enable Separate app sound and then set the following options - App > Choose an app to play its sound on a separate audio device. Audio device - Choose the audio device that you want the app's sound to be played on	Connect your device to a PC for an enhanced multitasking experience. Use your device and PC apps side by side. Share the keyboard mouse and screen between the two devices. Make phone calls or send texts while using DeX. Visit for more information - samsung.com/us/explore/dex
SQP(EWC + LRD)	From Settings tap Device care > Battery for options.	Connect to a Bluetooth device to make this option available in the Audio device menu. From Settings tap Sounds and vibration > Separate app sound. Tap Turn on now to enable Separate app sound and then set the following options	<SAME AS EMQAP>
SQP(SLR)	From Settings, tap Device care > Battery for options. Battery usage - View power usage by app and service. Power mode - Select a power life > App > power management. Configure Power.	From Settings tap and	<SAME AS EMQAP>
Remarks	For complex procedural questions, EMQAP give the answer closest to the ground truth.	For 'where' type questions, (asking the location of a particular feature), EMQAP again performs very well as compared to the other two variants.	Factual ('what' type) questions are answered equally well by EMQAP as well as the variants.

Table 4: Examples of question-answer pairs from the Samsung S10 QA Dataset and predictions by EMQAP and two variants (sentence-wise classification in AR model), with remarks, explaining the predictions.

MODEL	Factual	Procedural	Location
EMQAP	**0.455**	**0.582**	**0.664**
SQP(EWC+LRD)	0.417	0.576	0.561

Table 5: Average F1-Scores for factual, procedural and location-based questions on test set of S10 QA Dataset

Fig. 6 shows Ground Truth answers and the answers predicted by EMQAP and SQP(EWC+LRD) (both using sentence-wise classification) corresponding to three questions mentioned in Table 4. Fig. 7 similarly shows two more questions, but the first question shows how SQP(EWC+LRD) selects a wrong section when the answer is long, whereas, in the second question, EMQAP does not give the complete answer, while SQP(EWC+LRD) gives the correct answer.

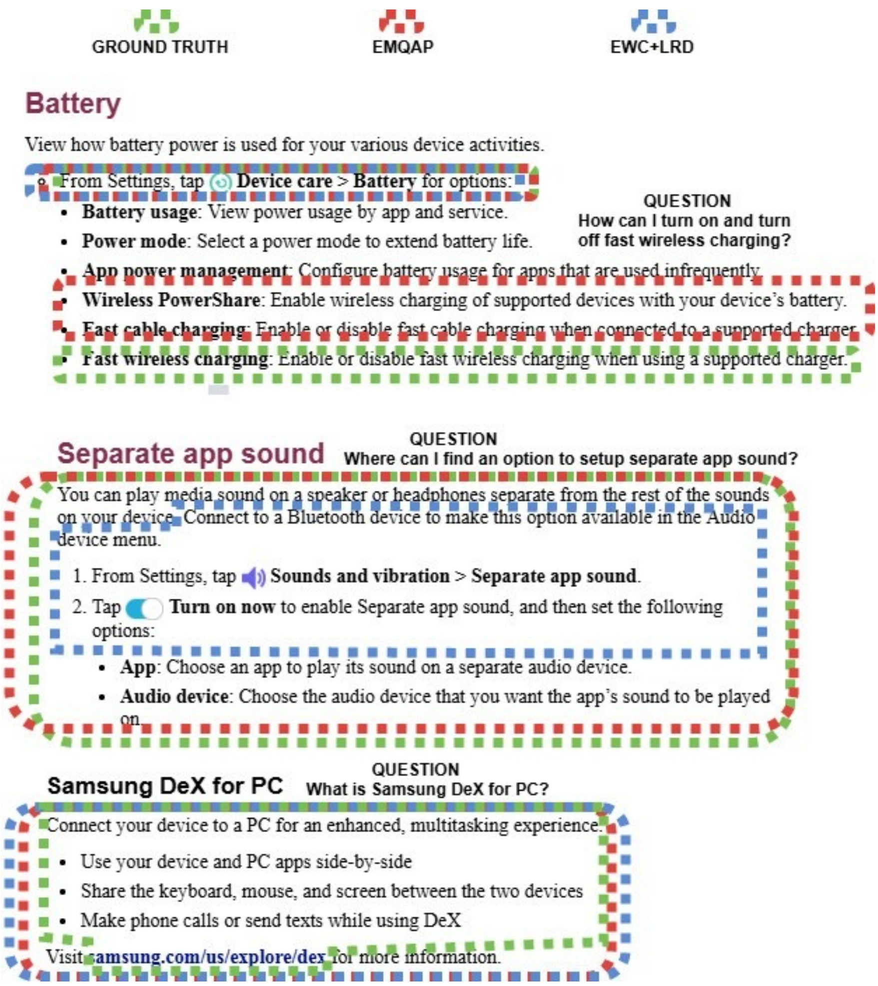

Figure 6: Ground Truth answers and the answers predicted by EMQAP and SQP(EWC+LRD) (both using sentence-wise classification) corresponding to three questions.

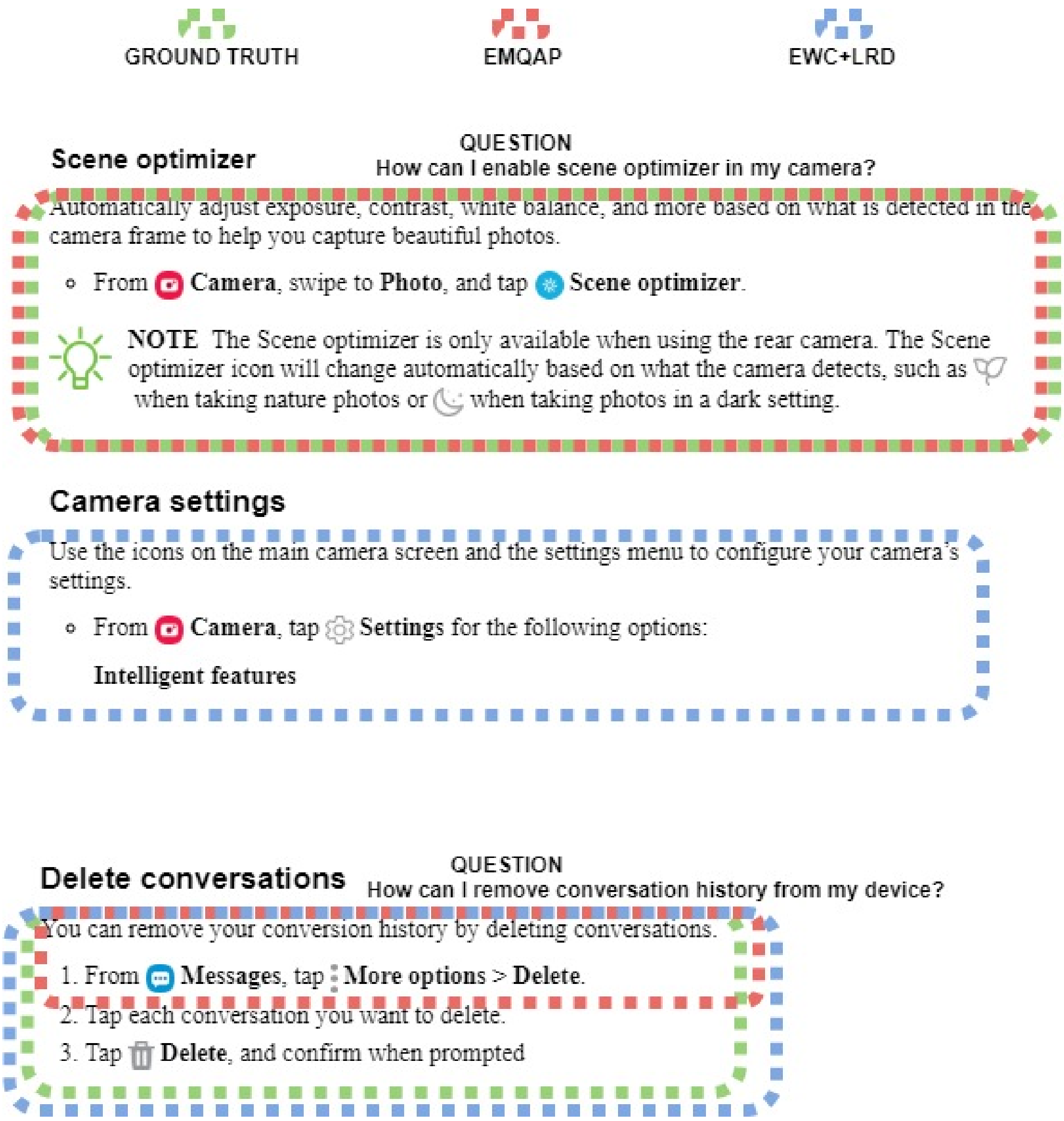

Figure 7: Ground Truth answers and the answers predicted by EMQAP and SQP(EWC+LRD) (both using sentence-wise classification) corresponding to two questions. In the first question, SQP(EWC+LRD) selects a wrong section, while in the second question, EMQAP does not give the complete answer.

5 Evaluating Smart TV annotated on CQA Forums

6 Evaluation on several devices

References

Kristjan Arumae, Qing Sun, and Parminder Bhatia. 2020. An empirical investigation towards efficient multi-domain language model pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4854–4864. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. In *Advances in Neural Information Processing Systems*, pages 13143–13152.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Aran Komatsuzaki. 2019. One epoch is all you need. *arXiv preprint arXiv:1906.06669*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *ArXiv*, abs/1904.08375.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexandre Rochette, Yadollah Yaghoobzadeh, and Timothy J Hazen. 2019. Unsupervised domain adaptation of contextual embeddings for low-resource duplicate question detection. *arXiv preprint arXiv:1911.02645*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.