Question Answering over Electronic Devices: A New Benchmark Dataset and a Multi-Task Learning based QA Framework (SUBMISSION ID: 4314)

Abhilash Nandy[♠] **Soumya Sharma**[♠] **Shubham Maddhashiya**[♠] **Kapil Sachdeva**[♣]
Pawan Goyal[♠] **Niloy Ganguly**[♠♢]
[♠]Indian Institute of Technology, Kharagpur [♣]Samsung Research Institute, Delhi
[♢] L3S Research Center, Leibniz Universität Hannover

1. Objectives

- Curation of a pre-training corpus of E-Manuals and an E-Manual QA Dataset
- Using different techniques of domain specific pre-training
- Leveraging Multi-Task Learning for Question Answering on E-Manuals

2. EMQAP Architecture

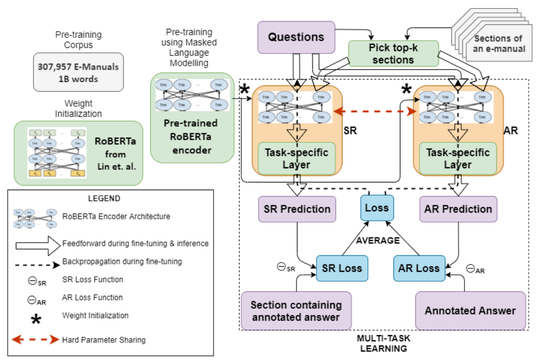

Figure 1: We propose EMQAP (E-Manual Question Answering Pipeline): RoBERTa architecture is used for pre-training on the corpus of E-Manuals, and its weights are used to initialize the MTL framework. A question with the topK relevant sections form inputs to the SR and AR modules of the MTL Framework during training, and an average of the AR and SR losses is backpropagated.

3. Dataset Collection and Extraction

- Large corpus of 3 lakh+ E-Manuals for domain-specific continual pre-training
- Annotated set of 900+ QA pairs for fine-tuning over MTL framework - Annotation to be done with help of domain-experts.
- Explore creating a community-based real QA dataset created using a question answering forum as an auxiliary dataset

4. Pre-training Corpus of E-Manuals

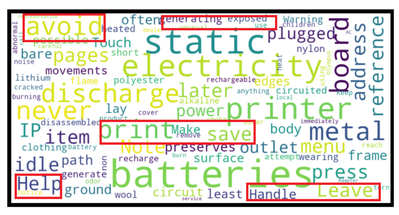

Figure 2: Instructional nature of the E-manuals corpus

Property	Value
No. of E-Manuals	307,957
No. of paragraphs	11,653,755
No. of sentences per paragraph	4.4
No. of words per sentence	20.2
Total number of words	∼1 Billion
Size of corpus (in GB)	∼11 GB

Figure 3: Details of the E-Manual pre-training corpus used in terms of property-value pairs

5. Domain-specific continual pre-training

Masked language Modelling with these variants -

- SLR (Same Learning Rate): pre-train RoBERTa [4] on E-Manuals with fixed Learning Rate across all layers
- LRD (Learning Rate Decay): pre-train RoBERTa on E-Manuals with Learning Rate decaying linearly across layers .
- EWC: pre-train RoBERTa on E-Manuals with Elastic Weight Consolidation (EWC) [3]
- EWC+LRD: Combination of EWC and LRD.

6. Annotated Question Answering Dataset

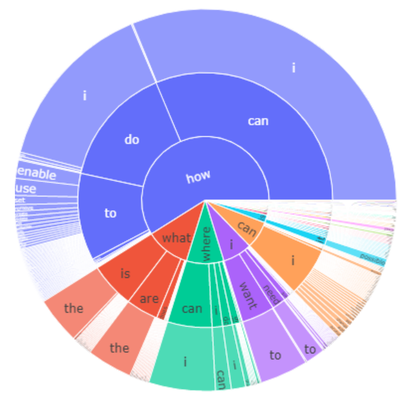

Figure 4: Distribution of questions covered in S10 QA Dataset w.r.t their first three tokens.

Dataset	Domain	No. of QA pairs	%age of factual questions	%age of procedural questions	%age of questions asking feature location	%age of paraphrased questions	Avg Question Length	Avg. Answer Length	Answer Type
TechQA (Castelli et al., 2020)	Technical Support	1,400	22.75	32.64	0.88	0	52.5	45	Single Span, long answer
S10 QA	E-Manual	904	7.08	48.34	7.3	33.52	9.4	48.4	Multi Span, long answer
Smart TV/Remote QA	E-Manual	950	14.26	51.74	3.03	30.35	11	61.5	Multi Span, long answer
Smart TV/Remote Amazon Consumer Questions	User Forum	1,028	12.35	37.06	0.97	0	12.84	20.41	

Figure 5: Description of our datasets and the TechQA Dataset. The % showing various categories (including the paraphrase) does not sum upto to 100 as some questions cannot be classified into one of the three categories.

7. Community-based QA Dataset

- A set of (Consumer Question, Annotated Question) pairs are annotated, such that each pair of questions are paraphrases of each other. This set can be used to train a Question Paraphrase Detector, which could extract questions from CQA Forum that are answerable using an E-manual.
- Corresponding to each such pair, we have two ground truths - one is the Annotated Ground Truth (AGT), and the other is the Consumer Ground Truth (CGT).

8. Baselines and comparative evaluation

- Method based on efficient passage retrieval Dense Passage Retrieval (DPR) [2]
- Methods with efficient answer retrieval - Technical Answer Prediction (TAP) [1] and MultiSpan [5]

MODEL	EM	P	R	F1	S+WMS
DPR	0	0.646	0.174	0.256	0.021
TAP	0.133	0.448	0.466	0.426	0.284
MultiSpan	0	**0.938**	0.14	0.226	0.014
EMQAP-T	0.156	0.577	0.682	0.588	0.34
EMQAP-S	**0.311**	0.801	**0.541**	**0.604**	**0.354**

Figure 6: Comparison of state-of-the-art models with EMQAP. (EMQAP-S and EMQAP-T are the Sentence-Wise and Token-Wise Classification variants, respectively

9. Evaluation on the Annotated Dataset

	Sentence-Wise Classification					Token-Wise Classification				
MODEL	EM	P	R	F1	S+WMS	EM	P	R	F1	S+WMS
SQP(FT RB)	0.178	0.696	0.457	0.506	0.273	0.133*	**0.59**	0.602	0.566	0.335
SQP(SLR)	0.156	0.733	0.473	0.522	0.246	0.033	0.587*	0.668	0.579	0.302
SQP(LRD)	0.256	0.783	0.507	0.57	0.321	0.089	0.559	0.603	0.539	0.295
SQP(EWC)	0.233	0.763	0.511	0.552	0.285	0.1	0.554	0.634	0.575	0.314
SQP(EWC+LRD)	0.278*	0.791*	0.523*	0.592*	0.33*	0.133*	0.574	0.673*	0.583*	0.337*
EMQAP	**0.311**	**0.801**	**0.541**	**0.604**	**0.354**	**0.156**	0.577	**0.682**	**0.588**	**0.34**

Figure 7: QA Evaluation on S10. "TF-IDF+T5" is applied by all the listed methods to select the top-10 relevant sections per question. EM stands for fraction of Exact Match. P(Precision), R(Recall) and F1 scores correspond to ROUGE-L. Best result for each metric is in **bold**, while the second best is marked with *

10. Evaluation on the Amazon CQA Dataset

GT	EM	P	R	F1	S+WMS
AGT	0.304	0.778	0.522	0.582	0.332
CGT	0.049	0.362	0.297	0.306	0.278

Figure 8: QA Evaluation on questions from CQA against corresponding answers from E-Manual of Samsung Smart TV as well as CQA. AGT is short for ANN-GT and CGT is short for CQA-GT ("TF-IDF+T5" is applied before all of the listed methods to select the top-10 relevant sections per question)

11. Evaluation on several devices

Sentence Wise Classification	Samsung Galaxy Mobile Phones	Other Samsung Mobile Phones	Samsung Tablets	Samsung Smart Watches
MTL (EMQAP)	**0.282**	**0.275**	**0.265**	**0.213**
SQP(EWC+LRD)	0.264	0.261	0.255	0.206

Figure 9: Average S+WMS scores on CQA Forum for 4 categories across 40 devices for EMQAP and variants, fine-tuned on S10 dataset. Best result for each category is in **bold**, while the second best is marked with *

12. Conclusion

In this paper, we worked on a far less studied problem of question answering from E-Manuals. In order to work the subject, a pre-condition was to create benchmark datasets which we painstakingly developed. We created a large corpus from E-manuals which was used in pre-training a RoBERTa architecture. This in turn helped in developing a domain-specific natural language understanding; the fruits of which can be observed in the huge improvement in performance over competing baselines. We believe that the E-manuals specific QA dataset is extensive and well-rounded and will help the community in various ways.

Acknowledgment: We would like to thank the annotators who made the curation of the datasets possible. Also, special thanks to Manav Kapadnis, an Undergraduate Student of Indian Institute of Technology Kharagpur, for his contribution towards the implementation of certain baselines. This work is supported in part by the Federal Ministry of Education and Research (BMBF), Germany under the project LeibnizKILabor (grant no. 01DD20003). This work is also supported in part by Confederation of Indian Industry (CII) and the Science & Engineering Research Board Department of Science & Technology Government of India (SERB) through the Prime Minister's Research Fellowship scheme. Finally, we acknowledge the funding received from Samsung Research Institute, Delhi for the work.
Email: nandyabhilash@gmail.com

References

[1] Vittorio Castelli, Rishav Chakravarti, Saswati Dana, Anthony Ferritto, Radu Florian, Martin Franz, Dinesh Garg, Dinesh Khandelwal, J. Scott McCarley, Mike McCawley, Mohamed Nasr, Lin Pan, Cezar Pendus, John F. Pitrelli, Saurabh Pujar, Salim Roukos, Andrzej Sakrajda, Avirup Sil, Rosario Uceda-Sosa, Todd Ward, and Rong Zhang. The techqa dataset. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of ACL 2020*, 2020.

[2] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*, pages 6769–6781, Online, November 2020.

[3] 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. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[5] Elad Segal, Avia Efrat, Mor Shoham, Amir Globerson, and Jonathan Berant. A simple and effective model for answering multi-span questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3074–3080, Online, November 2020. Association for Computational Linguistics.