# DellQA: developing a closed-domain QA model for IT support.

Cian Prendergast

21268862

31/07/2023

Word count: 2785

**Abstract**

This research proposes creating a high-quality custom QA dataset called DellQA for IT support. Data will be scraped from Dell's IT community support forum, and a domain expert will annotate question-answer pairs. A model already trained on SQuAD will be fine-tuned on the custom DellQA dataset. The study aims to address the scarcity of closed-domain QA datasets and demonstrate the feasibility of domain adaptation for IT support. DellQA's performance will be evaluated using standard metrics like EM and F1 score. The research highlights the importance of real-world questions from the support forum for better alignment with user queries.

**1. Introduction**:

IT support is a key element of a successful technology company, but it can be time-consuming and expensive to provide timely and helpful responses to so many customers and products (Yu et al., 2020). Dell Technologies has a vast community, answering questions posted on the community IT support forum. But it can be difficult to search and retrieve known solutions.

Question and Answering (QA) systems are a form of machine learning comprehension which extract relevant information (answers) from documents based on a prompt (question). But the initial requirement of a human annotator means that labelling open-domain question and answer pairs is easier than closed-domain, which requires expertise.

As a result there are only a small number publicly-available closed-domain QA datasets and only one IT support QA dataset called TechQA developed by IBM (Castelli et al 2019; Kia et al., 2022; Yu et al., 2020).

But Dell uses specific jargon not found in other technology companies like IBM. So even using this relevant closed-domain dataset (TechQA) will result in a poorly performing DellQA model (Soni & Roberts, 2020). So, this research proposes developing a small but high-quality custom QA dataset (DellQA). Data will be scraped via forum posts marked as solved and one domain-expert annotator will label the question-answer pairs. Then will follow the two-step process used by other researchers, taking a model pre-trained on a large open-domain dataset (SQuAD) and transfer-learn on a smaller closed-domain dataset (DellQA).

This research will contribute to the small number of IT support datasets and illustrate how companies can avail of readily available open-source tools to create their custom closed-domain QA datasets.

**2. Related Work**

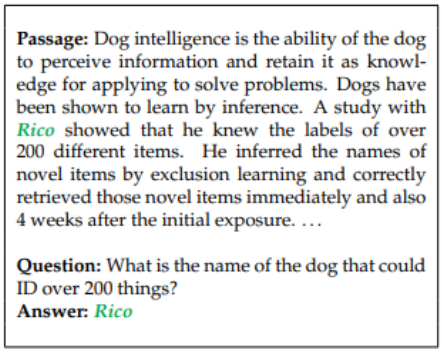
**2.1 Question & Answer:**

Question Answer (QA) systems are typically based on a retriever-reader architecture. First, a retriever pulls a relevant document (i.e., technical manuals) from a datastore. Dense passage retriever (DPR) is a specific dense retriever which use two Bidirectional Encoder Representations from Transformers (BERT) as encoders for questions and the answer passage (Tunstall, 2022). DPR's use of BERT helps when a user's phrasing or language used in the question doesn't exactly match up with the training set phrasing (Kia, 2022)**.** Next, to extract answers from the retrieved document we frame the problem as a span classification problem were the start and end tokens of an answer span act as the labels that a model needs to predict (Tunstall, 2022).

**2.2 SQuAD open-domain QA:**

To train QA models, a structured labelled dataset is required with one column containing the question, a second column containing the context (the entire text of the document), a column containing the span (start and end) of the answer present in the

context and finally the answer itself in text form (Tunstall, 2022). This dataset structure is derived from SQuAD (Rajpurkar et al, 2016) which consists of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage (Tunstall, 2022). Since the original SQuAD dataset, we have seen improvements with SQuAD 2.0 which contains questions designed to be unanswerable (Kwiatkowski, 2019). See figure 1 for an example of a passage (context), question and answer.


**Passage:** Dog intelligence is the ability of the dog to perceive information and retain it as knowledge for applying to solve problems. Dogs have been shown to learn by inference. A study with *Rico* showed that he knew the labels of over 200 different items. He inferred the names of novel items by exclusion learning and correctly retrieved those novel items immediately and also 4 weeks after the initial exposure. ...

**Question:** What is the name of the dog that could ID over 200 things?
**Answer:** *Rico*

*Figure 1, SQuAD question answer pair example (Soni & Roberts, 2020)*

However, QA models trained on large datasets like SQuAD, once deployed, often experience performance deterioration upon user-generated questions (Yu et al., 2020). This is because models trained on domain general corpus (Wikipedia) perform poorly with domain specific documents (Maufe, 2022). Thus, need closed-domain datasets which contain terminology/language unique to that specific domain (Kia et al., 2022).

**2.3 Closed-domain QA:**

Specialised closed-domain datasets remain scarce, with the most notable open-source examples residing in the medical domain (Hazen et al., 2019; Kia et al., 2022). This is because of the expense of getting a domain expert to spend time annotating answer spans is both time-consuming and costly. Open-domain QA datasets like SQuAD (Rajpurkar et al, 2016) can be crowdsourced for annotation, as they don't require domain expertise (Yogatama et al 2019).

This lack of domain specific datasets has contributed to closed-domain QA performing poorly compared to open-domain QA (Maufe, 2022). Some researchers have tried to overcome these issues of creating larger closed-domain QA datasets by using synthetic data generation (automatic question answer pairs) or by fine-tuning on unstructured source corpus (Bartolo et al., 2021; Yu et al., 2021). But these methods add complexity. Instead, by utilizing standard transfer learning techniques, we can take advantage of currently existing QA infrastructure and open-source libraries such as Hugging Face Transformers or deepset's Haystack. These tools reduce the barriers of implementation and makes repeatability by other researchers or technology companies easier (Cloudera Fast Forward Lab, 2020; Yu et al., 2020).

**2.4 Domain Adaptation:**

Thus, rather than increase the size of closed-domain QA datasets, Kia (2022) suggested a two-step fine-tuning process addresses the data scarcity problem for closed-domain QA: 1) transfer to the task; and 2) adapt to the target domain. Cloudera Fast Forward Lab (2020) found that closed-domain datasets can be small and effective, with no more than 2,000 question-answer pairs seeing good performance. Any greater saw diminishing returns on further performance gains.

A model trained on a large-scale SQuAD dataset learns the capability to answer questions in diverse domains and can quickly adapt to and perform efficiently in a new domain, as reflected in faster convergence and superior performance (Maufe, 2022; Hazen et al., 2019). This also suggests that smaller, closed-domain datasets must conform to the relatively complex SQuAD JSON format, as pre-trained models anticipate the input data to be in this specific structure (Tunstall, 2022)

**2.5 TechQA:**

TechQA is the only other publicly accessible closed-domain QA dataset that specifically targets IT support questions and answers (Yu et al., 2020). Data was collected from online IT support forums, where real users posed technical questions the community. A total of 276,968 forum threads were scraped from the website, filtered and then manually annotated by six people, one of which was a domain expert (Castelli et al 2019). The resulting set of question/answer pairs contains slightly more than 1000 items. The two-step process was used to train QA model: 1) BERT model pre-trained on the broad SQuAD dataset 2) transfer learning to domain specific TechQA (Kia, 2022).

SQUAD 2.0 (Rajpurkar et al, 2016) was designed to answer small factoid style questions while TechQA designed to answer technical questions. Thus, there was a length disparity with the initial TechQA dataset having an average question token length of 52 tokens compared to SQuADs 10 tokens (similar disparity for answers length). Castelli et al. (2019) opted to limit question-and-answer length to better align with SQuAD.

However, even with variation in answer lengths between datasets transfer learning is effective (Hazen et al., 2019). Also, a variation of BERT transformers called DistilBERT uses knowledge distillation and is able to capture features dependencies of long documents in a better way due to its triple loss combining (Alzubi et al., 2021).

**2.6 DellQA**:

The best QA model performance is obtained when the datasets are the same for both fine-tuning and prediction (Soni & Roberts, 2020). Thus, will develop a DellQA dataset to train and test a DellQA model.

DellQA model development will follow the two-step transfer learning process 1) DistilBERT language model fine-tuned on SQuAD, 2) Fine-tune on closed-domain custom dataset DellQA (Kia, 2022).

The DellQA performance will be evaluated using Exact Match (EM) and F1 score which is standard metrics for evaluating QA reader (Tunstall, 2022).

But the DellQA dataset will differ in its development compared to TechQA, data will be gathered from a data warehouse (using knowledge base articles marked as safe for public sharing) rather than scraping a website.

DellQA will not limit the token length of questions or answers, relying on DistilBERT's ability to handle longer questions/answers (Castelli et al 2019; Alzubi et al., 2021).

Also, Castelli et al. (2019) did not account for another alignment issue, the structure of the DellQA dataset must exactly match SQUAD (Tunstall, 2022).

**3. Data Collection**

**3.1 Scraping Dell Website (Data Collection)**

Dell hosts a community forum where individuals provide solutions to posted problems, many responders being Dell employees themselves. The website scraping focused on a subset 'Hardware' category of posts and where an answer given has been marked 'solved' by the community. A total of 1,901 question and answer pairs were collected from the website.

**3.2 Cleaning dataset**

Basic cleaning of dataset followed, removing meta tags, hashtags and signatures ect. But due to a limitation with Haystack Annotation tool, questions with more than 255 characters were truncated at the end. Answers had no character limit.

Data was exported as two csv files (questions and answers), following strict header guidelines and adding a document_identifier so they could be mapped together later.

### 3.3 SQUAD Formatting and labelling

The Haystack Annotation tool developed by deepset was used in order to 1) easily highlight the answer spans contained in community answers and 2) export in the complex SQuAD JSON format for later training.

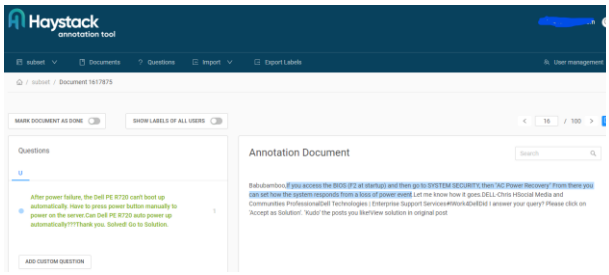



*Figure 2. Annotating answer spans.*

The two csv files were imported into the tool and paired using the document_identifier column. A total of 200 question and answer pairs were annotated using researcher's domain knowledge. The tool also allowed to mark questions as "don't understand" (i.e., if question was not in English) or "answer not given". Once labelled, dataset exported as a Json file which follows the structure of SQUAD with 180 answerable question-answer pairs.

### 3.4 Training

For tokenization a distilled version of BERT-base was used for its speed and the has been pre-trained on the larger SQUAD 2.0 dataset (downloaded from Hugging Face Hub).

In preparation for training, the dataset was randomly split into the train and test split following an 80:20 ratio.

Haystack was used to build the QA pipeline, which requires a document store for storing documents or passages for fast retrieval, a retriever responsible for finding relevant documents or passages in response to a given question, and a reader which extracts answer spans from those retrieved documents.

The DocumentStore used was Elasticsearch, chosen for its compatibility with dense retrievers and its ability to store and quickly search large volumes of text (Tunstall, 2022). A Dense Passage Retrieval (DPR) retriever was used to find information (passages) relevant to a query, leveraging its ability to handle variations in query phrasing. FARMReader was selected as the reader due to the ease of incorporating pre-trained models. All training was conducted within a Lambda Labs server.

## 4. Results

The primary aim of this research was to create a SQUAD dataset, but we also fine-tuned an existing model (trained on SQuAD) to the DellQA train set for one epoch and evaluated on our separate test set.

| | |
|---|---|
| EM | 27 |
| F1 | 44 |

*Figure 3. Fine-tuned for 1 epoch*

## 5. Discussion

QA is a critical tool for retrieving IT support information. People want solutions to problems and questions quickly. However, there are few publicly available closed-domain datasets– and only one currently for IT support (Yu et al., 2020). We want to contribute to the field by creating a second IT support QA dataset, DellQA.

This study used open-source tools to create a custom closed-domain dataset to train a QA model that performs well and brings business value. These tools are easily available and reduce the barriers for others to develop their own custom closed-domain QA.

We wanted to show that other companies can avail themselves of open-source tools to create small high quality structed datasets for efficient domain adaptation of custom QA models.

However, while Haystack annotation tool provides a GUI to label answer spans and exports automatically in SQuAD format, there

are some limitations. Firstly, the application requires strict adherence to standerdised headers (ie., document_indentifer) when importing as csv files. Secondly, the 255-character limit for questions results in truncating the ending of long-form questions, losing valuable information.

One of the significant benefits of sourcing the data from Dell support forums lies in the ability to acquire real-world questions, posed by actual users. This method provides a more authentic and representative dataset that is well-aligned with the types of inquiries the model would encounter in a practical setting.

Future research can expand the categories (not only 'Hardware') while scraping the Dell community forum. Also, using an alternative tool for annotation which allows for longer questions.

While only 200 question-answer pairs where annotated, this can easily be expanded using the raw dataset collected by scraping the support website.

**6. Conclusion**

In conclusion, this research addresses the scarcity of closed-domain QA datasets for IT support by introducing DellQA, a high-quality custom dataset. By scraping data from Dell's internal knowledge base and leveraging a domain expert's annotations, we successfully created DellQA. The fine-tuned DellQA model using a two-step transfer learning process demonstrates the feasibility of domain adaptation for IT support. Our study showcases the value of open-source tools like Hugging Face Transformers in facilitating the development of custom closed-domain QA datasets and models. The performance evaluation using standard metrics like EM and F1 score validates the effectiveness of DellQA. Overall, this work contributes to the advancement of closed-domain QA capabilities, enabling businesses to enhance their IT support services efficiently.

Supplementary materials, including Jupyter notebooks and the DellQA dataset associated with this study, are accessible here: https://github.com/c123ian/Dell_QA

**References**

Alzubi, J. A., Jain, R., Singh, A., Parwekar, P., & Gupta, M. (2021). Cobert: Covid-19 question answering system using bert. *Arabian Journal for Science and Engineering*. https://doi.org/10.1007/s13369-021-05810-5

Bartolo, M., Thrush, T., Jia, R., Riedel, S., Stenetorp, P., & Kiela, D. (2021). Improving question answering model robustness with synthetic adversarial data generation. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2021.emnlp-main.696

Castelli, V., Chakravarti, R., Dana, S., Ferritto, A., Florian, R., Franz, M., Garg, D., Khandelwal, D., McCarley, S., McCawley, M., Nasr, M., Pan, L., Pendus, C., Pitrelli, J., Pujar, S., Roukos, S., Sakrajda, A., Sil, A., Uceda-Sosa, R., … Zhang, R. (2020). The techqa dataset. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.117

Cloudera Fast Forward Lab. (2020, July 22). *Beyond SQuAD: How to apply a transformer QA model to your data*. NLP for Question Answering. Retrieved February 20, 2023, from https://qa.fastforwardlabs.com/domain%20adaptation/transfer%20learning/specialized%20datasets/qa/medical%20qa/2020/07/22/QA-for-Specialized-Data.html

Hazen, T. J., Dhuliawala, S., & Dhuliawala, D. (n.d.). Towards Domain Adaptation from Limited Data for Question Answering Using Deep Neural Networks. *Microsoft Research Montreal*. https://doi.org/10.48550/arXiv.1911.02655

Kia, M. A., Garifullina, A., Kern, M., Chamberlain, J., & Jameel, S. (2022). Adaptable closed-domain question answering using contextualized CNN-attention models and question expansion. *IEEE Access*, *10*, 45080–45092. https://doi.org/10.1109/access.2022.3170466

Kušniráková, D., Medved, M., & Horák, A. (2019). Question and answer classification in Czech question answering benchmark dataset. *Proceedings of the 11th International Conference on Agents and Artificial Intelligence*. https://doi.org/10.5220/0007396907010706

Maufe, M., Ravenscroft, J., Procter, R., & Liakata, M. (n.d.). A Pipeline for Generating, Annotating and Employing Synthetic Data for Real World Question Answering. *Research Gate*. https://doi.org/https://doi.org/10.48550/arXiv.2211.16971

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/d16-1264

Soni, S., & Roberts, K. (n.d.). Evaluation of Dataset Selection for Pre-Training and Fine-Tuning Transformer Language Models for Clinical Question Answering. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 5532–5538. Retrieved February 20, 2023, from https://aclanthology.org/2020.lrec-1.679.pdf.

Tunstall, L., Werra, L. von, & Wolf, T. (2022). Chapter 7, Question Answering. In *Natural language processing with transformers: Building language applications with hugging face*. essay, O'Reilly. Retrieved from https://huggingface.co/transformersbook.

Yogatama, D., de Masson d'Autume, C., Connor, J., Kocisky, T., Chrzanowski, M., Kong, L., Lazaridou, A., Ling, W., Yu, L., Dyer, C., & Blunsom, P. (2019). Learning and Evaluating General Linguistic Intelligence. *Deepmind*. https://doi.org/https://doi.org/10.48550/arXiv.1901.11373

Yu, W., Wu, L., Deng, Y., Mahindru, R., Zeng, Q., Guven, S., & Jiang, M. (2020). A technical question answering system with transfer learning. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. https://doi.org/10.18653/v1/2020.emnlp-demos.13

Yu, W., Wu, L., Deng, Y., Zeng, Q., Mahindru, R., Guven, S., & Jiang, M. (2021). Technical question answering across tasks and domains. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*. https://doi.org/10.18653/v1/2021.naacl-industry.23