The TechQA Dataset

Vittorio Castelli[1] **Rishav Chakravarti**[2] **Saswati Dana**[1] **Anthony Ferritto**[2]
Radu Florian[1] **Martin Franz** [1] **Dinesh Garg**[1] **Dinesh Khandelwal**[1]
Scott McCarley[1] **Michael McCawley**[3] **Mohamed Nasr**[1] **Lin Pan**[2]
Cezar Pendus[1] **John Pitrelli**[1] **Saurabh Pujar**[1] **Salim Roukos**[1]
Andrzej Sakrajda[1] **Avirup Sil**[1] **Rosario Uceda-Sosa**[1]
Todd Ward[1] **Rong Zhang**[1]

[1]IBM Research AI, [2]IBM Watson, [3] IBM Finance and Operations
{vittorio, rchakravarti, raduf, franzm, jsmc, mmccawley, mnasr, panl, cpendus, pitrelli,
roukos, ansa, avi, rosariou, toddward, zhangr}@us.ibm.com,
{aferritto, saurabh.pujar}@ibm.com, {sadana04, garg.dinesh, dikhand1}@in.ibm.com

Abstract

We introduce TECHQA, a domain-adaptation question answering dataset for the technical support domain. The TECHQA corpus highlights two real-world issues from the automated customer support domain. First, it contains actual questions posed by users on a technical forum, rather than questions generated specifically for a competition or a task. Second, it has a real-world size – 600 training, 310 dev, and 490 evaluation question/answer pairs – thus reflecting the cost of creating large labeled datasets with actual data. Hence, TECHQA is meant to stimulate research in domain adaptation rather than as a resource to build QA systems from scratch. TECHQA was obtained by crawling the IBMDeveloper and DeveloperWorks forums for questions with accepted answers provided in an IBM Technote—a technical document that addresses a specific technical issue. We also release a collection of the 801,998 Technotes available on the web as of April 4, 2019 as a companion resource that can be used to learn representations of the IT domain language.

1 Introduction

There is a tension between the development of novel capabilities in the early phases of the technology lifecycle, using unlimited data and compute power, and the later development of practical solutions as that technology matures. The challenges of creating practical solutions are twofold: developing robust, efficient algorithms and curating appropriate training data. Here we describe the curation and public release of a dataset intended to further those algorithmic advances.

The application domain is IT support, a notable component of the trillion-dollar IT services industry[1]. We created a dataset using publicly available data: questions from technical forums and answers from technical documents, all in English. We manually selected question-answer pairs that are appropriate for machine reading comprehension techniques, and reserved questions where the answer is distributed across multiple separate spans or documents, and those that require reasoning or substantial real world knowledge for future datasets. We release 600 questions for training purposes, of which 150 are not answerable from the provided documents, as well as 160 answerable and 150 non-answerable questions as development set. The blind test set contains 490 questions with similar answerable/non-answerable statistics to the development set.

The purpose of the TECHQA dataset is to stimulate transfer learning research from popular question-answering scenarios—driven by large-scale open-domain datasets with short questions and answers—to a use case with involved questions and often long answers. We expect that simple approaches based on tuning models trained on generic datasets will perform poorly on TECHQA, and that systems that are successful at the task embody algorithmic advances and novel approaches.

We are hosting a leaderboard for the TECHQA dataset at ibm.biz/Tech_QA where the data—training and development sets, as well as a collection of more than $800,000$ Technotes published on the internet—is available subject to registration. To maintain the integrity of the test set, the site provides the tools for authors evaluate their system on cloud infrastructure.

The rest of the paper is organized as follows.

[1]IT Service Report: https://www.selectusa.gov/software-and-information-technology-services-industry-united-states

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1269–1278

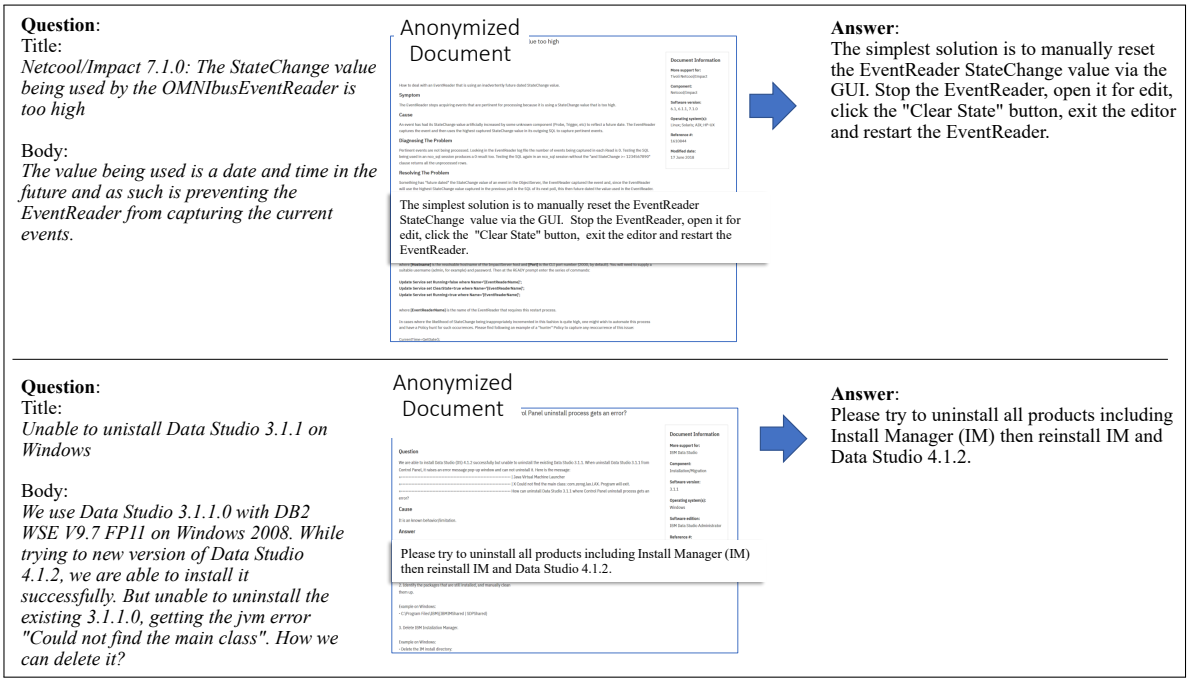

Figure 1: Examples of questions in the TechQA dataset.

We briefly review related work in Section 2; we then describe the process of collecting the data for TECHQA in Section 3, where we detail the automatic filtering, human filtering, annotation guidelines, and annotation procedure. We present statistics of the dataset in Section 4, introduce the associated leaderboard task in Section 5 and present baseline results obtained by fine-tuning MRC systems built for Natural Questions (hence-forth, NQ) (Kwiatkowski et al., 2019) and HOTPOTQA (Yang et al., 2018) in Section 6.

2 Related Work

Recent notable datasets for Machine Reading Comprehension (henceforth, MRC) include SQuAD 1.1 (Rajpurkar et al., 2016), SQuAD 2.0 (Rajpurkar et al., 2018), NarrativeQA (Kočiský et al., 2018) and HOTPOTQA. A common problem of the earlier MRC datasets is observation bias: annotators first read a paragraph and then wrote appropriate questions and answers, which, as a result, have substantial lexical overlap with the paragraph. Also, systems trained on SQuAD 1.1 could be easily fooled by the insertion of distractor sentences that should not change the answer, as shown in (Jia and Liang, 2017). Based on these considerations, SQuAD 2.0 added "unanswerable" questions. However, large pre-trained language models (Devlin et al., 2019; Liu et al., 2019) were able to achieve super-human performance in less than a year on SQuAD 2.0 as well; this suggests that the evidence needed to correctly identify unanswerable questions also are present as specific patterns in the paragraphs.

Recently, the NQ dataset has been introduced which overcomes the above problems and constitutes a much harder and realistic benchmark. The questions came from a commercial search engine and were asked by humans who had actual information needs. The answers were manually extracted from a Wikipedia page which the user may have selected among the search results.

HOTPOTQA is a recent multi-hop question-answering dataset (i.e., based on multi-step inference) where questions require reasoning over text from multiple Wikipedia pages. Systems must both produce answers and extract passages that contain supporting evidence.

All of the above datasets are said to be "open-domain", as the corpus is Wikipedia. There are also datasets for specialized domains. The biomedical QA dataset (Tsatsaronis et al., 2015) contained 29 development questions (arguably too few for training an automated system) and 282 test questions, divided into four categories–'yes/no', factoid, list, and summary. InsuranceQA (Ins), a dataset for the insurance industry, is a corpus for intent detection, rather than for MRC.

Our dataset, TECHQA, consists of questions posed in a technical forum by technical users who

Questions	Count
Total retrieved	276,968
With Accepted Answers	57,990
With link to Technote in Accepted Answer	15,918

Table 1: Statistics of questions from the forums. The questions with a Technote link in the Accepted Answer were manually annotated by our annotators.

had a specific information need, and answers from technical documents mentioned in the "Accepted Answer" to the post. In Section 4 we will contrast structural properties of TECHQA to those of some of the datasets mentioned here.

Datasets for specialized domain require effective domain adaptation (Wiese et al., 2017), because they contain a much smaller number of labeled examples than open-domain datasets like (Bajaj et al., 2016). Having a limited number of quality labeled examples is a real-world situation: domain experts are much more expensive than crowd-sourcing participants.

3 TECHQA Dataset Collection

The questions for the TECHQA dataset were posed by real users on public forums maintained and hosted by IBM at the developer.ibm.com answers[2] and IBM developerworks[3] sites. The questions are related to products running in environments supported by IBM and mostly fall into three categories: i) generic requests for information; ii) requests for information on *how to* perform specific operations; iii) questions about *causes of and solutions to* observed problems.

The questions are very specific: when describing an issue, the writer typically provides the versions of the affected software products, a description of the operations that yield the error, information about the error including portions of stack traces, and recent changes to the computing environment, such as upgrades, that might have bearing on the problem. Questions have a *title* and a *body*. The title is often an integral part of the question and therefore we include both title and body of the question in TECHQA.

As shown in Table 1, a significant fraction of the questions posted in the two forums have answers that were accepted by the person who asked the question (*accepted answers*). However, the majority of these Accepted Answers rely on the question or on fuller forum discourse history and are not good stand-alone candidates for a MRC dataset. For example *"You should be able to debug it – perhaps the value wasn't populated into that field when the messagebox was called."* is the accepted answer to the question *"how do I get the value of the dcedFirstName text field to display in my datacap custom verify panel?"*[4] Without context, this answer is uninformative, as are most of the answers in the forums.

About 6% of the accepted answers contain links to one or more Technotes, documents written and maintained by IBM support personnel that contain information about common questions asked by customers, product upgrade information, and official solutions to well-scoped problems. Technotes follow templates: for example, a troubleshooting Technote has an informative title, a description of the problem, an explanation of the cause, the products, versions, and configurations affected, steps to diagnose the problem, steps to solve the problem, and, if appropriate, temporary workarounds. Metadata in an infobox also describes the components, software version/editions, operating systems, and environments to which the Technote applies, as needed.

3.1 Automatic Filtering of Questions

The forums were crawled to return only those questions having the following characteristics: i) the question had an Accepted Answer; ii) the Accepted Answer contained a link to a Technote currently published on the web, and iii) the question was at most 12 sentences long. The last requirement was introduced because most question answering datasets described in Section 2 contain very short questions; since the goal of the TECHQA dataset is to promote domain adaptation, we opted to limit the question length for the TECHQA initial release.

We produced 15,918 candidate questions, which were manually annotated as described next.

3.2 Human Annotation

The candidate questions were reviewed by six annotators. Five are professional annotators with substantial experience in NLP annotation. The sixth is a Linux system administrator. Four annotators worked full time on the task while the other

[2] https://developer.ibm.com/answers/questions

[3] https://www.ibm.com/developerworks

[4] This question has been simplified and paraphrased in the interest of space.

two, including the system administrator, worked only part time. With the exception of the system administrator, who also acted in an advisory role, the annotators do not have a technical background.

Crucially, the annotators were not asked to answer technical questions, but to match the content of an Accepted Answer, provided by a subject matter expert in the forum, with the content of a technical document. To ensure that the annotators were comfortable with the subject matter of TECHQA, they were trained to annotate Technotes for mention detection according to an unreleased type system we developed for IT technical support and spent two months performing the mention detection task. When the TECHQA annotation started, they were familiar with the technical jargon and were able to read and understand both forums and Technotes. The annotators underwent a two-week training period on question and answers related to IBM products technical support, after which we annotated the TECHQA dataset. While generating TECHQA, we reviewed the results with the annotators twice a week to ensure quality and consistency of annotation.

3.3 Human Filtering of Questions

Question filtering consisted of inspecting question titles and bodies only, without considering the answers, and flagging questions that needed manual modification.

Some posts contain multiple questions in the question body. The prototypical case is a user reporting an error and asking for both cause of and solution to the problem. In some cases, the title and the body of the question appear to ask for different information as in:

- title: *"Where can I download the Integration Bus Healthcare Pack"*
- body: *"Where can I find information about the Integration Bus Healthcare Pack."*

When such questions were flagged by annotators, they were manually split into multiple separate questions each addressing a single information need, and re-submitted separately for annotation. We plan on releasing the unsplit questions in future releases of the dataset, where we will also allow answers consisting of separate spans from one or more documents.

The annotators also flagged questions to be manually modified as follows: i) stack traces embedded in questions were reduced by removing irrelevant information; ii) the signoff was removed when it contained a name; iii) product information available from parts of the forum other than the title and text of the questions was worked into the question text, if this modification was deemed necessary to make the question answerable. The original questions were disregarded and the modified questions resubmitted for annotation.

Only a small fraction of the questions were modified as a result of this and subsequent steps, constituting less than 10% of the released corpus, and most of the changes were very small.

3.4 Question-Answer Annotation Guidelines

The annotators were instructed to follow the guidelines for question selection and answer span selection outlined below.

3.4.1 Question Selection

Annotators were asked to identify the correct answer in the Technote linked from the forum accepted answer using question and Accepted Answer as guidance. Using question, accepted answer from the forum and Technote, the annotators were asked to discard questions that had the following characteristics:

i) The Accepted Answer in the forum is excessively long (longer than 10 sentences). We do this because annotators found long Accepted Answers difficult to match with the content of the Technote. It was left to annotators' discretion to retain long accepted answers whenever they felt that the information was clear.

ii) The answer in the Technote is excessively long. Answers exceeding 10 sentences should be discarded.

iii) The Technote does not contain an answer to the question. This happens when the Accepted Answer points to Technotes that are topical but not essential to the answer. For example, the answer might state that the product mentioned in the question is an old version that should be updated before addressing the problem and points to a Technote describing the update process.

iv) The answer consists of multiple separate spans of text. Future releases of the dataset will address domain adaptation for multi-hop question-answering systems.

v) The answer is distributed across multiple Technotes.

3.4.2 Answer Span Selection

As a result of discussions with IBM subject matter experts, we instituted the following guidelines for answer span selection. The annotators were instructed to select the shortest span that would answer a question for an expert in the field. The annotators were also asked to select the answer to the specific question asked in the forum, and not to add topical information to the answer span: if the post asks for the cause of a problem, the answer should not include the solution; conversely, the answer to a post about solving a problem should not contain information about the cause.

Text surrounding the actual answer and containing information already provided in the question must not be included in the answer. For example, consider the problem of upgrading a component under Windows® 10 and a Technote that lists the steps for various OS. The sentence *"These are the steps for Windows® 10"* should not be part of the selected answer. Similarly, examples are not deemed to be part of the answer unless they are short and occur in the middle of the answer.

3.5 Annotation and Adjudication

Each question that passed the automatic filtering and manual filtering was independently annotated by two annotators.

Questions that were selected by at least one annotator were further manually adjudicated. The two authors who oversaw the annotation reviewed disjoint subsets of the annotator results and were allowed to perform the following operations:

- select the answer of one of the annotators when the two annotators disagreed;
- reduce the span of the answer, while conforming to the directives listed above;
- flag a question as containing multiple questions, when both annotators failed to recognize it;
- shorten the question, mostly by removing parts of stack traces (a process that could be easily automated);
- occasionally reject the answer—by-and-large when one of the annotators had already rejected the answer.

The two authors who supervised the annotation task also independently annotated 100 answerable questions; the inter-annotator agreement F1 is 76.3% and the exact match rate is 61%.

The resulting set of question/answer pairs released with the dataset contains slightly more than 850 answerable questions, and slightly fewer than 550 non-answerable questions. In future versions of the TECHQA, we plan to relax many of these annotator constraints to promote research addressing a broader spectrum of tech support problems.

4 TECHQA Dataset Characteristics

The TECHQA dataset consists of a training set, a development set, a test set, and a small validation set. The training set contains 450 answerable questions and 150 non-answerable questions, the development set consists of 160 answerable and 150 non-answerable questions, and the evaluation set consists of 490 questions with similar answerable vs. non-answerable ratio as the development set. The ratios of non-answerable to answerable questions in the splits are similar to those of SQuAD 2.0 (Rajpurkar et al., 2018). The validation set consists of the first 20 entries of the development set and is used in the leaderboard described in Section 5. We also provide the full collection of the unique 801,998 Technotes that were available on the web as of April 4, 2019.

The dataset is designed for MRC, rather than for open-domain QA. Specifically, instead of requiring users to search the Technote collection to find one containing the answer, we provide for each question a candidate list of 50 Technote IDs. Systems should analyze only the 50 Technotes associated with the question. A question is answerable if the annotators found an answer in one of these 50 Technotes, and is unanswerable otherwise. Systems can access the entire Technote collection but only answers from the 50 Technotes associated with each questions will be scored. The 50 Technotes were obtained by issuing a query to an instance of Elasticsearch[5] that indexes the 801,998 Technotes. This query consisted of the concatenation of the question title and question text; thus, the retrieved Technotes are expected to contain at least some of the low-frequency terms in the question. If the answer is in a Technote not retrieved by the search engine, we randomly removed one of the 50 Technotes and substituted it with the one containing the answer. We did not include the search engine scores of the Technotes and we randomized their order to obfuscate their search engine ranking.

[5] https://www.elastic.co/products/elasticsearch

TECHQA questions and answers are substantially longer than those found in common datasets. Table 2 compares statistics of training and development sets questions and answers of TECHQA to those of SQuAD 2.0 and HOTPOTQA, in white-space-separated tokens. Figures 2 and 3 depict the length distributions for questions text, title plus text, and answers for training and devset, respectively. Most questions have a length between 10 and 75 tokens, but the dataset exhibits a long tail, reflecting the fact that questions with a substantial amount of detailed information are relatively common. Most answers are between 1 and 100 tokens long, and the distribution has a long tail. A typical question consists of a description of the issue experienced by the person who posted it, while the actual "question" is typically short, as illustrated by the second example of Figure 1, where the question is "How we can delete it?" [sic].

The questions and answers contain numerous technical terms. We estimated the number of mentions of technical support entities with a model built with the mention detection data produced by our annotators during their training period (see Section 3.2). On average the training set questions contain 1.67 detected mentions of errors, error codes, error messages or log messages (we do not further extract mentions from error messages or log messages, hence the subsequent counts are from other parts of the question), 3.8 mentions of hardware or software products or components, 2.0 mentions of parameters, settings, or configurations, and 2.23 mentions of operations or specific commands issued by the person asking the question, among others. Many of these terms are likely not part of the vocabulary of most general-purpose contextual language models. Hence, one of the reasons for including the whole Technotes corpus is to provide data for enhancing the language models by appropriately enlarging the vocabulary to include technical support terms.

5 Leaderboard task

The dataset is available by registering to the leaderboard at ibm.biz/Tech_QA. Registered users have access to the data and to means for submitting systems for evaluation against the blind test set. As with other leaderboards, this approach will help maintain the integrity of the blind set.

A submitted system must be packaged as a Docker image, containing all the needed components. The container will run in isolation from the network: systems will not be allowed to download anything–including models or other resources–while running in the evaluation environment. The systems will read the evaluation data from a read-only input directory and will write results to an output directory. Detailed instructions on how to package the system are available from the leaderboard site. We ask that systems submitted to the leaderboard do not use information from the developer.ibm.com answers[6] and IBM developerworks[7] web sites except for the data provided with the dataset.

Submitted systems will run on a machine with 128 GB of memory and two 16G V100 GPUs, with 64 GB local disk space available for temporary files or logs. Upon submission, the system will run against the 20-question validation set. The results of the validation run are made available on the user's personal dashboard. A user satisfied with the validation run can submit the system to be run against the 490 evaluation questions. Runs will be limited to 24 hours, after which they will be terminated and the submission will be in an error state in the dashboard. Successful runs are added to the dashboard.

The user can monitor the progress of each submission from the dashboard, and cancel the submission at any point previous to completion of the evaluation run. The results of successful evaluation runs are automatically posted on the leaderboard. A user is prevented from submitting a new system for a week starting from the date of the most recent submission, as it appears on the public leaderboard. The user dashboard provides means for anonymizing and de-anonymizing a successful submission (for example, for paper review purposes). An anonymized submission retains the name of the system provided by the user, but hides the user's affiliation as well as the optional link to a paper.

Systems are required to analyze the 50 documents associated with each question, and produce 5 candidate answers. Each answer consists of a document ID, start and end character offsets from the beginning of the detagged text of the Technote, and a score. The score is compared with a threshold provided by the system for the run. Systems must return scores lower than the threshold

[6] https://developer.ibm.com/answers/questions
[7] https://www.ibm.com/developerworks

Dataset	Split	Question length in tokens				Split	Answer length in tokens			
		min	mean	max	std		min	mean	max	std
SQuAD 2.0	training	1	9.9	40	3.4	training	1	3.2	43	3.4
	devset	3	10.0	31	3.45	devset	1	3.1	29	3.1
HotpotQA	training	3	17.8	108	9.5	training	1	2.2	89	1.8
	devset	6	15.7	46	5.5	devset	1	2.5	29	1.8
TechQA	training	8	52.1	259	31.6	training	1	48.1	302	37.8
	devset	10	53.1	194	30.4	devset	1	41.2	137	27.7

Table 2: Statistics of the question and answer lengths in white-space-separated tokens for SQuAD 2.0, HotpotQA and TechQA.

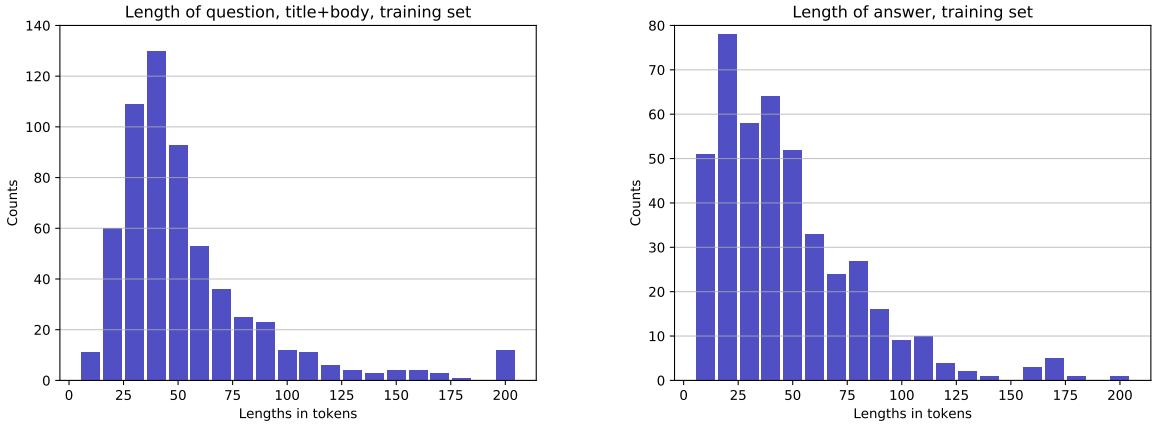

Figure 2: Number of white space separated tokens in training questions (title plus body.) and answers (for answerable questions only). The bin at 200 also contains all questions longer than 200 tokens.

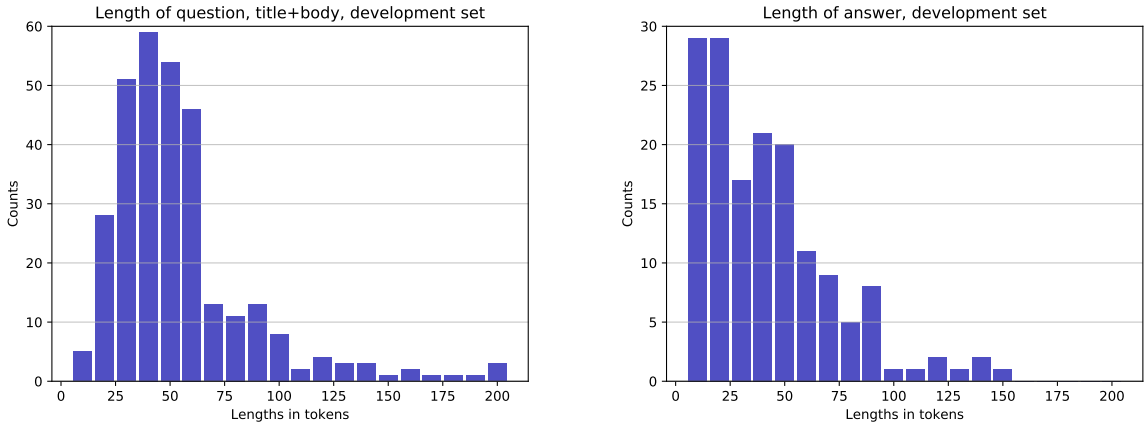

Figure 3: Number of white space separated tokens in devtest questions (title plus body) and answers (for answerable questions only).

to indicate that no answer exists in the Technote; however, they also must indicate the best span extracted from the document: this is used to compute the two ancillary metrics described below.

The evaluation score computed for the leaderboard is a zero-one value for a question/document pair with score below the threshold, and character-overlap F1 for a question/document pair with score greater than or equal to the threshold.

The main metric, called F1 on the leaderboard, is the macro average of the evaluation scores computed on the first of the five answers provided by the system in response to each question.

The leaderboard displays three ancillary metrics. HA_F1@1 is the macro average of the evaluation scores computed on the first of the five answers and averaged over the answerable evaluation questions. This metric should be compared to the inter-annotator agreement of 76.3 reported in Section 3. HA_F1@5 consists of computing the evaluation score for each of the 5 answers, selecting the maximum, and computing the macro average over all answerable questions. BEST_F1 is the value of the F1 metric corresponding to the optimal choice of the threshold. The time required for the run will also be made available.

6 Baseline Results

Table 3 show the results of three baseline systems on the development set. These are a model trained on SQuAD 2.0, a model trained on NQ, and the TAP system submitted to HOTPOTQA[8].

Both SQuAD and NQ models consists of a $\text{BERT}_{\text{LARGE}}$ (whole word masking) language model (Devlin et al., 2019) with additional layers. For SQuAD 2.0 these are two fully connected FF layers followed by softmax for answer begin- and end-boundary extraction, like in (Devlin et al., 2019). The NQ model further adds a layer for target type prediction as in (Alberti et al., 2019), tuned as described in (Pan et al., 2019). The table contains entries for both models out-of-the box and after fine-tuning on the TECHQA dataset.

The TAP system consists of a document ranker module followed by an answer span selector, both based on pretrained BERT small. If the largest score produced by the ranker exceeds a threshold, the question is declared answerable and the answer span selection is invoked on the documents.

[8]https://hotpotqa.github.io/

Table 3 shows that, without domain adaptation, the SQuAD and NQ models fail to produce interesting answers, and their best performance is roughly that of a dumb system that declares all questions unanswerable. Fine-tuning yields a notable improvement for both models. The TAP model has slightly lower performance but yields the highest HA_F1@5.

7 Discussion and Future Work

We have introduced TECHQA, a question-answering dataset for the IT technical support domain. The overall size of the released data (600 training questions) is in line with real-world scenarios, where the high cost of domain expert time limits the amount of quality data that can reasonably be collected. Thus, the dataset is meant to stimulate research in domain adaptation, in addition to developing algorithms for longer questions and answers than the current leaderboards.

We have created a leaderboard to evaluate systems against a blind dataset of 490 questions with a ratio of answerable to unanswerable questions similar to that of the development set. The leaderboard ranks submissions according to a metric consisting of the character overlap F1 measure for answerable questions and the zero-one metric for non-answerable questions. The leaderboard also reports the F1 at the top result and at the top 5 results averaged over the answerable questions.

TECHQA is a challenging dataset for models developed for existing open-domain MRC systems. Their out-of-the box performance is very low, especially considering that a system that declares every question as unanswerable achieves F1=48.4% on the development set. The obvious approach of fine-tuning these models using the TECHQA training set yields systems that barely beat the baseline.

The initial version of the dataset was created by selecting questions and answers that are relevant to the IT technical support domain but at the same time do not diverge excessively from the spirit of other existing MRC datasets. We consider TECHQA to be a stepping stone on which to build future data collections and leaderboards.

We plan on releasing questions with answers in a broader and more diverse collection that will include documents with a less formulaic structure than the Technotes. We will also relax the length limitations to include questions rich in details, and

Systems	F1	HA_F1@1	HA_F1@5	BEST_F1
SQuAD 2.0 − FT	1.67	3.25	4.51	48.39
SQuAD 2.0 + FT	54.05*	22.01	35.50	54.05
NQ − FT	2.74	5.32	9.07	48.39
NQ + FT	**55.31***	**34.69**	50.52	**55.31**
TAP_v0.1	**51.36**	16.39	**57.49**	52.67

Table 3: Baseline systems performance the dev set. The first 4 systems were pre-trained on the dataset indicated in the first column. In the same column, '−FT' indicates no fine-tuning after pre-training, while '+FT' indicates further fine-tuning using the TECHQA corpus. Entries marked with '∗' use a threshold tuned on the development set using the F1 metric; hence, F1 equals BEST_F1.

answers that include complex procedures; in the same spirit, we will allow answers consisting of multiple spans from a single document.

Many answers cannot be obtained by extracting portions of a document based on language alone: in many cases, domain knowledge is needed and often a question cannot be answered from the data collection without reasoning steps. We envision a roadmap where future releases of TECHQA will require synergy between multiple AI disciplines, from deep-learning based MRC to reasoning, knowledge base acquisition, and causality detection.

Acknowledgments

Our gratitude goes to our annotators: Abraham Mathews (IBM), Kat Harkavy, Irina Paegelow, Daniele Rosso, Chie Ugumori and Eva Maria Wolfe (ManpowerGroup Associates), for their dedication to TECHQA and their relentless effort.

References

InsuranceQA, a question answering corpus in insurance domain. https://github.com/shuzi/insuranceQA. Last commit: January 16, 2017.

Chris Alberti, Kenton Lee, and Michael Collins. 2019. A BERT baseline for the natural questions. *arXiv preprint arXiv:1901.08634*, pages 1–4.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

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

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *TACL*, 6:317–328.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a benchmark for question answering research. *TACL*.

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. *CoRR*, abs/1907.11692.

Lin Pan, Rishav Chakravarti, Anthony Ferritto, Michael Glass, Alfio Gliozzo, Salim Roukos, Radu Florian, and Avirup Sil. 2019. Frustratingly easy natural question answering.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. *EMNLP*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq

large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017. Neural domain adaptation for biomedical question answering. *CoNLL*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.