Team JARS: DialDoc Subtask 1 - Improved Knowledge Identification with Supervised Out-of-Domain Pretraining

Sopan Khosla, Justin Lovelace, Ritam Dutt, Adithya Pratapa
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA
{sopank,jlovelac,rdutt,vpratapa}@andrew.cmu.edu

Abstract

In this paper, we discuss our submission for DialDoc subtask 1. The subtask requires systems to extract knowledge from FAQ-type documents vital to reply to a user's query in a conversational setting. We experiment with pretraining a BERT-based question-answering model on different QA datasets from MRQA, as well as conversational QA datasets like CoQA and QuAC. Our results show that models pretrained on CoQA and QuAC perform better than their counterparts that are pretrained on MRQA datasets. Our results also indicate that adding more pretraining data does not necessarily result in improved performance. Our final model, which is an ensemble of AlBERT-XL pretrained on CoQA and QuAC independently, with the chosen answer having the highest average probability score, achieves an F1-Score of 70.9% on the official test-set.

1 Introduction

Question Answering (QA) involves constructing an answer for a given question in either an extractive or an abstractive manner. QA systems are central to other Natural Language Processing (NLP) applications like search engines, and dialogue. Recently, QA based solutions have also been proposed to evaluate factuality (Wang et al., 2020) and faithfulness (Durmus et al., 2020) of abstractive summarization systems.

In addition to popular QA benchmarks like SQuAD (Rajpurkar et al., 2016), and MRQA-2019 (Fisch et al., 2019), we have seen QA challenges that require reasoning over human dialogue. Some notable examples being QuAC (Choi et al., 2018) and CoQA (Reddy et al., 2019). These datasets require the model to attend to the entire dialogue context in the process of retrieving an answer. In this work, we are interesting in building a QA system to help with human dialogue.

Feng et al. (2020) introduced a new dataset of goal-oriented dialogues (Doc2Dial) that are grounded in the associated documents. Each sample in the dataset consists of an information-seeking conversation between a user and an agent where agent's responses are grounded in FAQ-like webpages. DialDoc shared task derives its training data from the Doc2Dial dataset and proposes two subtasks which require the participants to (1) identify the grounding knowledge in form of document span for the next agent turn; and (2) generate the next agent response in natural language.

In this paper, we describe our solution to the subtask 1. This subtask is formulated as a span selection problem. Therefore, we leverage a transformer-based extractive question-answering model (Devlin et al., 2019; Lan et al., 2019) to extract the relevant spans from the document. We pretrain our model on different QA datasets like SQuAD, different subsets of MRQA-2019 training set, and conversational QA datasets like CoQA and QuAC. We find that models pretrained on out-of-domain QA datasets substantially outperform the baseline. Our experiments suggest that conversational QA datasets are more useful than MRQA-2019 data or its subsets. In the following sections, we first present an overview of the DialDoc shared task (§2), followed by our system description (§3) and a detailed account of our experimental results, and ablation studies (§4, §5).

2 DialDoc Shared Task Dataset

Dataset used in the DialDoc shared-task is derived from Doc2Dial dataset (Feng et al., 2020), a new dataset with goal-oriented document-grounded dialogue. It includes a set of documents and conversations between a user and an agent grounded in the associated document. The authors provide annotations for dialogue acts for each utterance in the

Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 103–108
August 5–6, 2021. ©2021 Association for Computational Linguistics

dialogue flow, along with the span in the document that acts as the reference of it.

The dataset shared during the shared task was divided into train/validation/testdev/test splits. Train and validation splits were provided to the participants to facilitate model development. During phase 1, the models were evaluated on testdev whereas, the final ranking was done on the performance on the test set.

Pre-processing Using the pre-processing scripts provided by the task organizers, we converted the Doc2Dial dataset into SQuAD v2.0 format with questions containing the latest user utterance as well as all previous turns in the conversation. This is in line with previous work from (Feng et al., 2020) which showed that including the entire conversational history performs better than just considering the current user utterance. Dialogue context is concatenated with the latest user utterance in the reverse time order.

The output of this pre-processing step consisted of 20431 training, 3972 validation, 727 testdev, and 2824 test instances.

3 System Description

As discussed earlier, subtask 1 of DialDoc shared task is formulated as a span selection problem. Therefore, in order to learn to predict the correct span, we use an extractive question-answering setup.

3.1 Question-Answering Model

We pass the pre-processed training data through a QA model that leverages a transformer encoder to contextually represent the question (dialogue history) along with the context (document). Since the grounding document is often longer than the maximum input sequence length for transformers, we follow (Feng et al., 2020) and truncate the documents in sliding windows with a stride. The document trunk and the dialogue history are passed through the transformer encoder to create contextual representations for each token in the input. To extract the beginning and the ending positions of the answer span within the document, the encoded embeddings are sent to a linear layer to output two logits that correspond to the probability of the position being the start and end position of the answer span. The training loss is computed using the Cross-Entropy loss function. We use the huggingface transformers toolkit in all of our experiments.

3.2 Pretraining

Recent work (Gururangan et al., 2020) has shown that multi-phase domain adaptive pretraining of transformer-based encoders on related datasets (and tasks) benefits the overall performance of the model on the downstream task. Motivated by this, we experimented with further pretraining the QA model on different out-of-domain QA datasets to gauge its benefits on Doc2Dial (Table 1).

QA Dataset	Domain	# Samples
SQuAD	Wikipedia	86k
NewsQA	News	74k
NaturalQuestions	Wikipedia	104k
HotpotQA	Wikipedia	73k
SearchQA	Jeopardy	117k
TriviaQA	Trivia	62k
MRQA-19 (Train)	Mixed	516k
QuAC	Wikipedia	70k
CoQA	Kids' Stories, Literature, Exams, News, Wikipedia	70k

Table 1: Statistics (domain, # samples) for different QA datasets used for continual pre-training.

4 Experimental Setup

In this section, we discuss our experimental setup in detail.

4.1 Pretraining Datasets

Firstly, we briefly describe the different datasets used for the continual pretraining of our transformer-based QA models.

MRQA-19 Shared task (Fisch et al., 2019) focused on evaluating the generalizability of QA systems. They developed a training set that combined examples from 6 different QA datasets and developed evaluation splits using 12 other QA datasets. We explored the effectiveness of pretraining on the entire MRQA training set as well on each of the 6 training datasets: **SQuAD** (Rajpurkar et al., 2016), **NewsQA** (Trischler et al., 2017), **NaturalQuestions** (Kwiatkowski et al., 2019), **HotpotQA** (Yang et al., 2018), **SearchQA** (Dunn et al., 2017), and **TriviaQA** (Joshi et al., 2017).

Conversational QA datasets. We also experiment with pretraining on two conversational QA datasets: **QuAC** (Choi et al., 2018)[1] and

[1] https://huggingface.co/datasets/quac

QA Dataset	Validation EM	Validation F1
Doc2Dial	42.1	57.8
+ SQuAD	45.0	**60.3**
+ NewsQA	**45.5**	59.8
+ NaturalQuestions (NQ)	44.2	59.9
+ HotpotQA	43.0	58.0
+ SearchQA	42.3	57.5
+ TriviaQA	43.1	58.0
+ MRQA-19 (Train)	43.4	58.9
+ SQuAD + NewsQA + NQ	43.0	59.2
+ SQuAD + NewsQA + NQ (IS)	43.8	59.4
+ QuAC	46.4	60.3
+ CoQA	**47.7**	**66.0**

Table 2: Performance (EM (%), F1 (%)) of `bert-base-uncased` on DialDoc validation set when further pretrained on different QA datasets.

CoQA (Reddy et al., 2019).[2] For both datasets, we filter out samples which do not adhere to SQuAD-like extractive QA setup (e.g. yes/no questions) or have a context length of more than 5000 characters.

Table 1 presents the size of the different pretraining datasets after the removal of non-extractive QA samples.

4.2 Evaluation Metrics

The shared-task relies on Exact Match (EM) and F1 metrics to evaluate the systems on subtask 1. To compute these scores, we use the metrics for SQuAD from huggingface.[3]

4.3 Hyperparameters

We use default parameters set by the subtask baseline provided by the authors.[4] However, we reduce the training per-device batch-size to 2 to accommodate the large models on an Nvidia Geforce GTX 1080 Ti 12GB GPU. We stop the continual out-of-domain supervised pretraining after 2 epochs.

5 Results

We now present the results for different experimental setups we tried for DialDoc subtask 1.

5.1 Pretraining on Different QA Datasets

Our first set of results portray the differential benefits of different out-of-domain QA datasets when used to pretrain the transformer encoder.

QA Dataset	Validation EM	Validation F1	Testdev EM	Testdev F1	Test EM	Test F1
`bert-large-uncased-whole-word-masking`						
Doc2Dial	50.1	63.4	–	–	–	–
+ SQuAD	52.4	63.9	–	–	–	–
+ QuAC **(1)**	53.2	68.0	47.4	66.5	–	–
+ CoQA **(2)**	54.3	70.3	49.4	**68.7**	45.5	65.5
+ CoQA,QuAC **(3)**	54.2	70.1	**51.0**	68.1	–	–
`albert-xl`						
+ QuAC **(4)**	59.1	72.6	47.6	67.1	52.6	67.4
+ CoQA **(5)**	60.0	74.1	48.0	67.9	50.8	69.5
`Ensembles`						
E (4,5)	61.4	75.3	49.5	66.6	**53.5**	**70.9**
E (1,2,3,4,5)	**61.5**	**76.1**	49.5	**68.7**	52.0	69.9

Table 3: Performance (EM (%), F1 (%)) of large transformer-based QA models on DialDoc validation and testdev set when further pretrained on different QA datasets. Scores in **bold** are the best in their column; **underlined** are best on the official test-set.

Experiments with `bert-base-uncased` on the validation set (Table 2) portray that pretraining on different QA datasets is indeed beneficial. Datasets like SQuAD, NewsQA, and NaturalQuestions are more useful than SearchQA, and TriviaQA. However, pretraining on complete MRQA-2019 training set does not outperform the individual datasets suggesting that merely introducing more pretraining data might not result in improved performance. Furthermore, conversational QA datasets like CoQA and QuAC, which are more similar in their setup to DialDoc, perform substantially better than any of the other MRQA-2019 training datasets.

We observe similar trends with larger transformers (Table 3). Models pretrained on QuAC or CoQA outperform those pretrained on SQuAD. However, combining CoQA and QuAC during pretraining does not seem to help with the performance on validation or testdev split.

Analyzing Different Transformer Variants Table 3 also contains the results for experiments where `albert-xl` is used to encode the question-context pair. We find that `albert-xl`-based models outperform their `bert` counterparts on validation set. However, they do not generalize well to the Testdev set, which contains about 30% of the test instances but is much smaller than validation set in size (727 samples in testdev vs 3972 in validation set).

[2] https://huggingface.co/datasets/coqa

[3] https://huggingface.co/metrics/squad

[4] https://github.com/doc2dial/sharedtask-dialdoc2021/

5.2 Results on test set

We only submitted our best performing models on the official test set due to a constraint on the number of submissions. Contrary to the trends for testdev phase, `albert-xl` models trained on conversational QA datasets perform the best. `albert-xl + QuAC` is the best-performing single model according to the EM metric ($EM = 52.60$), whereas `albert-xl + CoQA` performs the best on F1 metric ($F1 = 69.48$) on the test set.

5.3 Ensembling

We perform ensembling over the outputs of the model variants to obtain a single unified ranked list. For a given question Q, we produce 20 candidate spans, along with a corresponding probability score ps. We compute rank-scores rs for the answer-spans at rank r as $rs = \frac{1}{\log_2(r+1)}$. We then aggregate the information of the answer spans for the model variants using the following techniques.

Frequent: We chose the answer span which was the most frequent across the model variants.

Rank Score : We chose the answer span which was the highest average rank score.

Probability Score: We chose the answer span which was the highest average probability score.

We observe empirically that ensembling using the probability score performs the best and hence we report the results of ensembling using the probability score (**E**) in Table 3.

We observe the highest gains after ensembling the outputs of all the 5 model variants on the validation test and test-dev set. However, the best performance on the test set was achieved by ensembling over the `albert-xl` models pre-trained independently on CoQA and QuAC (EM = 53.5, $F1$ = 70.9). This was the final submission for our team.

5.4 Informed Data Selection

We investigate the disparate impact of pretraining on different MRQA-19 datasets on the Doc2Dial shared task. Specifically, we explored factors such as answer length, relative position of the answer in the context, question length, and context length in Table 4. We observe that the SQuAD, NewsQA, and NaturalQuestions (NQ) has comparatively longer answers than the other datasets. However, we do not observe a noticeable difference in terms of question length, context length or relative position of the answer in the context, with respect to the other datasets.

Dataset	Question	Answer	Context	Rel Pos
SQuAD	59.6	20.2	754.7	0.462
NaturalQ	47.2	23.7	804.8	0.390
NewsQA	36.8	25.0	3022.7	0.261
TriviaQA	76.1	9.7	4069.3	0.380
SearchQA	80.4	10.9	3818.7	0.392
HotpotQA	114.0	14.3	945.0	0.457
Doc2Dial	61.4	129.3	4814.2	0.427

Table 4: Statistics of Average Question Length, Average Answer Length, Average Context Length, and Average Relative Position of the Answer in the Context for Doc2Dial and different MRQA subsets.

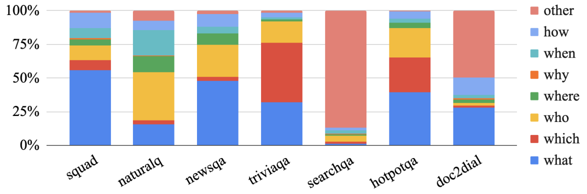

Figure 1: Distribution of Question Words for MRQA.

We also use the dataset of Li and Roth (2002) to train a BERT classifier to predict answer type of a question with 97% accuracy. The coarse-answer types are DESC (Description), NUM (Numerical), ENT (Entity), HUM (Person), LOC (Location) and ABBR (Abbreviation). We use the classifier to gauge the distribution of answer types on MRQA datasets and Doc2Dial. We observe from Figure 2 that a majority of questions in Doc2Dial require a descriptive answer. These DESC type questions are more prevelant in SQuAD, NewsQA, and NQ, which might explain their efficacy.

To ascertain the benefit of intelligent sampling, we pretrain on a much smaller subset of the SQuAD, NewsQA, and NaturalQuestions dataset, which we obtain via intelligent sampling. We select questions which satisfy one of the following criteria, (i) the answer length of the question is ≥ 50, (ii) the question includes 'how' or 'why' question word or (iii) the answer type of the question is 'DESC'. Overall, the size of the selected sample is only 20% of the original dataset, yet achieves a higher EM score than the combined dataset as seen in Table 2. Yet, surprisingly, the performance is lower than each of the individual dataset.

6 Conclusion

Our submission to the DialDoc subtask 1 performs continual pretraining of a transformer-based encoder on out-of-domain QA datasets. Experiments

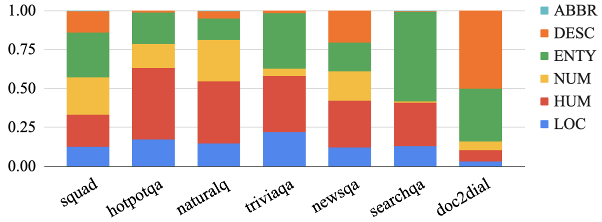

Figure 2: Distribution of Answer Types for MRQA.

with different QA datasets suggest that conversational QA datasets like CoQA and QuAC are highly beneficial as their setup is substantially similar to Doc2Dial, the downstream dataset of interest. Our final submission ensembles two AlBERT-XL models independently pretrained on CoQA and QuAC and achieves an F1-Score of 70.9% and EM-Score of 53.5% on the competition test-set.

Impact Statement

In this work, we tackle the task of question answering (QA) for English language text. While we believe that the proposed methods can be effective in other languages, we leave this exploration for future work. We also acknowledge that QA systems suffer from bias (Li et al., 2020), which often lead to unintended real-world consequences. For the purpose of the shared task, we focused solely on the modeling techniques, but a study of model bias in our systems is necessary.

References

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Dunn, Levent Sagun, Mike Higgins, V. U. Güney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new qa dataset augmented with context from a search engine. *ArXiv*, abs/1704.05179.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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. *Transactions of the Association of Computational Linguistics*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. Association for Computational Linguistics.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

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*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.