Conversational QA for FAQs

Jon Ander Campos[1], **Arantxa Otegi**[1], **Aitor Soroa**[1],
Jan Deriu[2], **Mark Cieliebak**[2], **Eneko Agirre**[1]
[1]University of the Basque Country (UPV/EHU)
[2]Zurich University of Applied Sciences (ZHAW)
[1] {jonander.campos, arantza.otegi, e.agirre, a.soroa}@ehu.eus
[2]{jan.deriu, mark.cieliebak}@zhaw.ch

Abstract

The goal of this work is to access the large body of domain-specific information in the form of Frequently Asked Question sites via conversational Question Answering (QA) systems. Training systems for each possible application domain is unfeasible, calling for research on transfer learning of conversational QA systems. We present DoQA, a dataset for accessing **Do**main specific FAQs via conversational **QA** that contains 1,637 information-seeking dialogues on the cooking domain (7,329 questions in total). These dialogues are created by crowd workers that play the following two roles: the **user** who asks questions about a certain cooking topic posted in Stack Exchange, and the **domain expert** who replies to the questions by selecting a short span of text from the long textual reply in the original post. The expert can rephrase the selected span, in order to make it look more natural. Together with the dataset, we present results of state-of-the-art models, including transfer learning from Wikipedia QA datasets to our cooking FAQ dataset, and a more realistic scenario where the passage with the answer needs to be retrieved. Our dataset and experiments show that it is possible to access domain specific FAQs with high quality using conversational QA systems with little training data, thanks to transfer learning.

1 Introduction

Access to textual information via search has its limitations, as it returns documents for each search query, which need to be examined by the user. With the raise in popularity of textual chat- and voice-mediated access to information, the user expects the system to reply to inter-connected questions with short, up-to-the-point answers. Given an incomplete or partially correct answer, the user likes to ask follow-up questions, giving the system additional opportunities to satisfy the user needs. The overarching objective of our work is to build the technology to access the large body of domain-specific information in the form of Frequently Asked Question sites (FAQ for short) via conversational Question Answering (QA) systems. Note that we include in the generic concept of FAQs also Community Question Answering sites, as well as corporate information in intranets which is maintained in textual form similar to FAQs, often referred to as internal "knowledge bases".

More specifically, in this paper we present **DoQA**, a task and associated dataset for accessing Domain Specific FAQs via conversational QA, together with the evaluation of existing conversational QA systems. The dataset contains 1,637 information-seeking question/answer dialogues on the cooking domain (7,329 questions in total). These dialogues are created by crowd workers that play the following two roles: the **user** asks questions about a given topic posted in Stack Exchange[1], and the **domain expert** replies to the questions by selecting a short span of text from the long textual

[1]`https://stackexchange.com/`

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

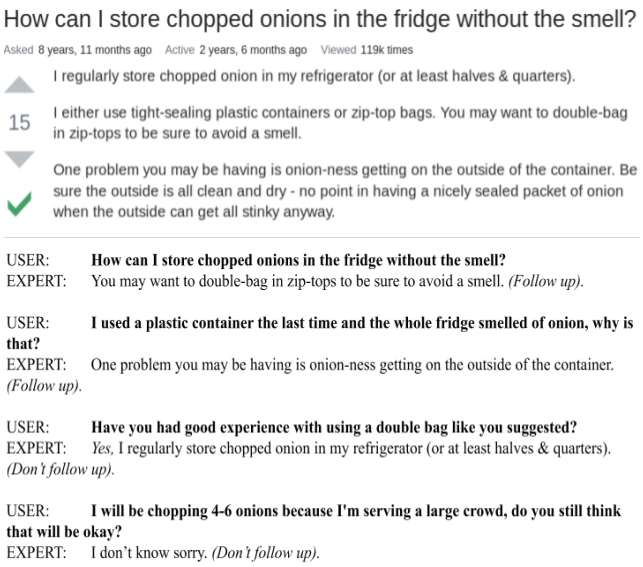

Figure 1: An example dialogue about a Stack Exchange cooking topic. On top, the original post, comprising a topic and long answer. Below, the collected dialogue. The user, who only knows about the topic, asks free form questions. The expert, based on the post answer, provides answers.

reply in the original post. We focused on the cooking domain[2], as it is one of the most active, and contains knowledge of general interest, making it easily accessible for crowd workers. In addition to the selected span, we also allow experts to rephrase it, in order to provide a more natural answer. DoQA enables the development and evaluation of conversational QA systems that help users access the knowledge buried in domain specific FAQs.

Current technology to access FAQs is limited to a single-turn answer, in the form of a snippet of the target document produced by query-focused summarization techniques. More recently, conversational QA datasets like CoQA (Reddy et al., 2018) and QuAC (Choi et al., 2018) have collected large numbers of human conversational QA dialogues, showing that collecting such datasets is feasible. Still, those datasets cannot be used to train systems for our problem. In general, FAQs contain open-ended non-factoid questions with complex and subtle replies. On contrast, the questions of CoQA were produced with access to the target document and answers are very short, and QuAC is limited to Wikipedia articles about people.

Together with the dataset, we present results of existing state-of-the-art conversational QA models, including transfer learning from Wikipedia QA datasets to our cooking FAQ dataset. An information retrieval module is also evaluated, in order to provide results in a more realistic scenario for conversational QA. Our dataset and experiments show that it is possible to produce high quality conversational QA systems on specific domains using with little training data, thanks to transfer learning. The gap with respect to human performance shows that there is ample room for system improvement.

2 Related work

Conversational QA systems stem from the body of work on Reading Comprehension, whose goal is to test the capacity of a system to understand a document by answering any question posed over its content. Recent work on the field has resulted in the creation of multiple datasets (Rajpurkar et al., 2016; Trischler et al., 2017; Nguyen et al., 2016; Kočiský et al., 2018; Dunn et al., 2017). These datasets are typically composed of multiple question/answer pairs, often along with a reference passage from which the answer is curated.

More similar to our work, CoQA (Reddy et al., 2018) and QuAC (Choi et al., 2018) are two conversational QA datasets comprising QA dialogues that fulfill the information need of an user by answering questions about different topics. Similarly to our, both datasets are built by crowdsourcing,

[2] `https://cooking.stackexchange.com/`

where one person (the questioner) is presented with a topic and has to pose free-form questions about it. Another person (the answerer) has to select an answer to the question by choosing an excerpt from the relevant passage describing the topic. Some of the questions in both datasets are unanswerable, and context about the topic is needed in order to answer some of the questions.

CoQA contains 127k questions with answers, obtained from 8k conversations about passages from broad domains, ranging from children stories to science. The answers are also excerpts from the relevant passage, but answerers have the choice of reformulating them. The authors report that 78% of the answers had at least one edit. Although reformulating answers can yield to more natural dialogues, Yatskar (2018) showed that span based systems can in principle obtain a performance up to 97.8 points F1, showing that editing the answers does not yield to systems with better quality. In CoQA, both questioner and answerer have access to the full passage, which greatly guides the conversation towards the specific information conveyed in it. This strategy is similar to the one used in the CoQA dataset (Reddy et al., 2018).

In contrast, in our dataset DoQA the passage is not shown to the person asking questions, and thus the questions in the dialogue rely on his/her own intuition and information needs. In addition, the questions in CoQA are specific, many times about factoids, and the answers tend to be very short. Please see Section 4 for a head-to-head comparison.

QuAC is a dataset that contains 14k information-seeking question answering dialogues. The dialogues in QuAC are about a specific section in Wikipedia articles about people. The answerer has access to the full section text, whereas the questioner only sees the section's title and the first paragraph of the main article, which serves as in inspiration when formulating the queries. QuAC also contains dialogue acts in each turn, which are useful when collecting the dialogues, as they can be used by the answerer to indicate to questioner whether to continue making questions about the last answer or drift to other aspects of the topic. The questions in QuAC also tend to be factoids about people, while DoQA focuses on open-ended questions about specific topics. Please see Section 4 for a head-to-head comparison.

In conversational QA datasets the relevant document or passage that contain the answer of a query is provided, which greatly facilitates the task of the system. However, in a real world scenario the queries must be answered by searching over big information sources such as Wikipedia or the whole web. Chen et al. (2017) and Watanabe et al. (2017) combine retrieval and answer extraction on a large set of documents to answer the question. In (Talmor and Berant, 2018) the authors propose decomposing complex questions into a sequence of simple questions, and using search engines to answer the single question, from which the final answer is computed.

In DoQA we include a ranking of multiple documents that can be relevant to answerer a query and hence it may be used to assess the ability of conversational QA systems to perform a dialog when the exact passage with the correct answers are not known.

3 Dataset collection

This section describes our cooking conversational QA dataset [3] collection process which consists of an interactive task designed for two crowd-workers in Amazon Mechanical Turk (AMT).

3.1 AMT task

We define a HIT as the task of generating a dialogue about cooking between two workers. As said before, one of the workers (the **user**) asks questions to the second one (the **domain expert**) about a certain topic from a Stack Exchange[4] cooking thread. The worker who adopts the **user** role has access to a small paragraph that introduces the topic. Having this information, he must ask free text questions. The first question of every dialogue must be the title of the topic that appears in the title of the Stack Exchange thread.

The **domain expert** has access to the whole answer passage and he/she answers the query by selecting a span of text from it. In order to make the dialogue look more natural, the domain expert has the opportunity to edit the answer, but note that if he does so the answer will not match the content of the

[3]The DoQA dataset is available here: `http://ixa2.si.ehu.es/convai/doqa-v1.0.zip`

[4]We downloaded the data dump from September 2018.

Retrieval	MAP		P@1		R@20	
	dev	test	dev	test	dev	test
Question	0.97	0.95	0.97	0.94	0.99	0.98
Answer	0.73	0.65	0.65	0.54	0.92	0.88

Table 1: Results of the question and answer retrieval for dev and test using two possible strategies. Mean Average Precision (MAP), Precision at 1 (P@1) and Recall at 20 (R@20) are given.

text span anymore. Therefore, and following (Yatskar, 2018), we motivate minimal modifications by copying the selected text span directly into the answer field in the web application. In addition to the span of text, the expert has to give feedback with one of the following dialogue acts:

- Continuation. It is used for leading the user to the most interesting topics: *follow up* or *don't follow up*.
- Affirmation. It is required when the question is a Yes/No question: *yes*, *no* or *neither*.
- Answerability. It will define if the question has an answer or not: *answerable* or *no answer*. When no answer is selected, the returned string is "I don't know".

These dialogue acts are the same as in QuAC, but we discarded the *maybe follow up* act from the continuation set because we feel that it is not very intuitive.

Dialogues are ended when a maximum of 8 question and answer pairs is reached, when 3 unanswerable questions have been asked, or when 10 minutes time limit is reached. The purpose of these limits is to avoid long and repetitive dialogues, because real cooking threads are very focused on a certain topic, and usually there are very few long and repetitive conversations.

3.2 Dataset details

Following usual practice, we divided the dataset into a train, development and test splits, with 1037, 200 and 400 dialogues respectively.

In the test split we do not allow more than one dialogue about the same section, as it can end up producing inaccurate evaluation of the models.

3.3 Collecting multiple answers

In order to estimate the performance of a human in the task, we collected additional answers after having completed the dialogues in the test split. This is also useful for evaluation, as some of the questions in DoQA can have more than one valid answer. In this addtional collection, a single worker had to provide an answer span for each question in the test split. All previous questions and answers of the respective dialogue were also shown to the worker, so he is aware of the previous dialogue history. We use these additional answers as a way of measuring how hard the questions are, as difficult questions are expected to have multiple and diverse answers.

3.4 Information retrieval scenario

In the usual setting for this kind of tasks, the system is given the question and the passage where the answer is to be extracted from. In a realistic scenario, however, relevant answer passages that may contain the answer will need to be retrieved first. More specifically, if a user has an information need and asks a question to a conversational QA system on a FAQ, the system can search for similar questions which have already been answered, or the system can directly search in existing answer passages.

Table 1 shows the results of the question and answer retrieval approaches. The question retrieval approach yielded very good results as expected, because the questioner workers of the AMT task most of the time started the dialogue asking the same question that was posted in the forum, even if they often edited and rewrote it. Thus, most of the input queries for the IR system where equal to the ones that were indexed. The results section shows the results of the conversational QA system when relying on the passages returned by the IR module.

Dataset	DoQA	QuAC	CoQA
Questions	7,320	98,407	**127,000**
Dialogues	1,637	**13,594**	8,399
Tokens / question	**10.68**	6.5	5.5
Tokens / answer	13.03	**14.6**	2.7
Questions / dialogue	4.47	7.2	**15.2**
Extractive %	68.71	**100**	66.8
Abstractive %	31.29	-	**33.2**
Yes/No %	20.81	**25.8**	-
I don't know %	**28.06**	20.3	1.3

Table 2: Statistics of DoQA compared to QuAC and CoQA.

Bigram prefix	%	Example
What	16.6	
is	30.8	What is the purpose of adding water to an egg wash?
are	8.0	What are other methods to sharpen a knife?
How	15.1	
do	24.0	How do you properly defrost frozen fish?
long	21.9	How long should I cook it in the microwave?
Is	10.5	
there	52.8	Is there a special tool available for cracking open a pistachio?
it	19.8	Is it safe to cook with rainwater?
Do	7.6	
you	70.7	Do you have any advice for storing green onions?
I	16.1	Do I have to peel the apples?
Can	5.5	
I	52.8	Can I put them back in the oven to reheat?
you	25.3	Can you explain the science behind this cooking procedure?

Table 3: The most frequent initial words and phrases of the questions of DoQA.

4 Dataset analysis

In this section we present an quantitative and qualitative analysis of DoQA and we compare them to similar conversational datasets like QuAC and CoQA, stressing its similarities and differences.

Overall statistics Table 2 shows the overall statistics of DoQA, together with the statistics of QuAC and CoQA. As can be seen, DoQA has the smallest amount of questions and dialogues. However, other features makes it very interesting for the research of conversational QA. For instance, the average tokens per questions and answers (10.68 and 13.03, respectively) are closer to real dialogues if we compare to the other datasets. Specially CoQA has very short questions and answers on average, as they are only 5.5 and 2.7 words long, suggesting that CoQA is closer to factoid QA than dialogue, as human dialogues tend to be longer and convoluted, not just short answers. DoQA has the lower ratio of questions per dialogue, which is expected, as most of the dialogues are about a very specific topic and the user is satisfied and gets the answer without the need of long dialogues. Regarding to the percentages of extractive and abstractive answers, they are similar to the ones of CoQA, suggesting that in a similar way to Yatskar (2018) we could develop both robust extractive and abstractive systems for our developed dataset. QuAC lacks this abstractive feature. With respect to dialogue acts, DoQA is similar to QuAC. CoQA suffers from the lack of dialogue acts and ends up on having almost all of its questions answerable, facing the same issues as SQuAD 1.0 (Rajpurkar et al., 2016) that motivated the addition of unanwerable questions in SQuAD 2.0 (Rajpurkar et al., 2018).

Question types Table 3 provides the most frequent two initial words of the questions in DoQA along with their percentages of occurrences and some examples. These figures and examples give an insight into the types of questions in DoQA. Most of the questions start with *what* and *how* (16.6% and 15.1% of the questions, respectively), which are also the most frequent in QuAC and CoQA. Contrary to them the questions in DoQA do not refer to factoids, with the exception of "How long

questions". The questions in DoQA, as exemplified in the table, require long and complex answers. In contrast to this, in CoQA and QuAC many of the most frequent initial wordssuch as *who*, *where*, and *when* indicate factoid questions. In order to confirm this fact, we manually inspected 100 random questions, and we could see that more than half of the questions are non-factoid in DoQA, showing that most of the questions are open-ended.

Context or history dependence The manual analysis also shows that 61% of the questions are dependent on the conversation history, as many questions have coreferences to previous questions or answers in the dialogue. Some of the examples in Table 3 show this phenomenom: *What are other methods to sharpen a knife?*, *How long should I cook it in the microwave?*, *Can you explain the science behind this cooking procedure?*. Moreover, we could note that less than 1% ask further advice or tips about the current topic, confirming that these conversations are about specific topics where the user is satisfied with the expert answers after a few questions.

5 Task definition

Given a textual passage and the a question, traditional QA systems find an answer to the question within the passage. Conversational QA systems are more complex, as they need to deal with a sequence of possibly inter-dependent questions. That is, the meaning of the current question may depend on the dialogue history. For this reason, a dialogue history comprised by previous question/answer pairs is also provided to the system. In addition, some dialogue acts have to be predicted as an output: yes/no answers, which are required for affirmation questions, and continuation feedback, which might be useful for information-seeking dialogues.

We denote the answer passage as p, the dialogue history of questions and respective ground truth answers as $\{q_1, a_1, ...q_{k-1}, a_{k-1}\}$, current question as q_k, the answer span a_k which is delimited by its starting index i and ending index j in the passage p, and dialogue act list v. The dialogue act list contains *{yes,no,-}* values for predicting affirmation and *{follow-up,don't follow-up}* for continuation feedback.

6 Baseline models

We implemented two strong baseline models to address the above task.

BERT and BERT+Yes/No This baseline is an adaptation of BERT, which has shown strong performance on QA datasets such as SQuAD (Devlin et al., 2018). We took the fine-tuning approach for QA of BERT as a starting point, which already predicts the indexes i and j of the a_k answer span given p and q_k as input. In addition, we modified this model to get a version that is able to also predict the list v of dialogue acts in addition to the answer text span. Our approach to predict dialogue acts relies on the final hidden vector of the `[CLS]` embedding.

We modeled the prediction of affirmation and continuation feedback dialogue acts as follows: we added a classification layer $W \in \mathbb{R}^{K \times H}$ to the last hidden vector of the `[CLS]` token, followed by a softmax. K denotes number of labels and H hidden size. The former model is referred to as BERT, and the later as BERT+Yes/No.

BERT+HAE The previous baseline does not model dialogue history. We used BERT with History Answer Embedding (Qu et al., 2019) as a baseline that deals with the multi-turn problem, as this is the publicly available system that performs best in the QuAC leaderboard [5]. The system introduces dialogue history $\{q_1, a_1, ...q_{k-1}, a_{k-1}\}$ to BERT by adding a history answer embedding layer, which learns whether a token is part of history or not.

7 Evaluation

Evaluation metrics Given the similarity between QuAC and DoQA, we use the same evaluation metrics and criteria used in QuAC. F1 is the main evaluation metric and is computed by the overlap

[5] accessed on August 20, 2019

Model	F1	HEQ-Q	HEQ-D	Y/N	F1-all	F1	HEQ-Q	HEQ-D	Y/N	F1-all
BERT	–	27.00	0.5	–	35.93	41.4	38.6	4.8	-	36.2
BERT+Yes/No	–	25.5	0.5	76.9	33.9	40.2	35.4	6.2	**78.0**	36.1
BERT+HAE	–	27.66	1.0	–	40.72	**47.8**	**43.0**	**7.8**	-	**42.7**
Human						86.7				

Table 4: Results (dev on left side, test on right side) of the baseline models (and human performance) trained and tested on DoQA.

Model	Fine-tune	F1	HEQ-Q	HEQ-D	Y/N	F1-all	F1	HEQ-Q	HEQ-D	Y/N	F1-all
BERT	-	-	29.10	1.0	-	34.13	41.3	36.2	4.8	-	37.6
BERT+Yes/No	-	-	27.2	0.5	75.6	33.3	41.9	39.0	4.2	77.1	36.6
BERT+HAE	-	-	30.66	0.5	-	40.34	46.2	42.0	6.5	-	42.3
BERT	DoQA	-	28.2	1.0	-	40.6	44.7	41.2	7.5	-	40.9
BERT+Yes/No	DoQA	-	28.9	1.0	78.9	38.6	46.5	42.0	7.8	**80.0**	41.7
BERT+ HAE	DoQA	-	36.56	1.5	-	46.53	**54.6**	**50.3**	**10.8**	-	**48.4**

Table 5: Results of the baseline models following the transfer learning approach. All the experiments are trained on QuAC and tested on DoQA. The last three lines shows the results when the model is fine-tuned using DoQA train. The differences between the rows are the model and the data used for fine tuning.

at word level of the prediction and reference answers. As the test set contains multiple answers for each question we take the maximum F1 among them. We also report HEQ (human equivalence score) which measures the percentage of examples for which system F1 exceeds or matches human F1, with two variants: HEQ-Q, which is computed on a question level, and HEQ-D, which is computed on a dialogue level. For dialogue acts of affirmation and continuation feedback, we report accuracy with respect to the majority annotation. Note that when computing F1 QuAC filers out answers with a low agreement among human annotators. An additional F1-all is provided for the whole answer set.

Experimental setup We first carried out experiments using the extractive information of the train/dev/test splits of DoQA. The parameters we used for baseline models training are the ones proposed in the original papers. In this case we use the train split for training the BERT model and the dev test for early stopping. Then, following the transfer learning approach, we used the train data of QuAC for training the BERT model. Once having this new model we tested it directly on DoQA test split. Moreover, we analysed the benefits of fine tuning this last model trained on QuAC with the DoQA train split.

We also experiment using the provided IR rankings, which contain the top 20 passages for each dialogue. In the first experiment, dubbed "Top-1", we just use the top 1 passage in the BERT+Yes/No model. In a second experiment, dubbed "Top-20 / BERTprobs", the passages are fed to the BERT+Yes/No model and the passage that contains the answer with highest confidence score is selected. Note that we discard passages that produced "I don't know" type of answers. In a third experiment, dubbed "Top-20 /BERTprobs & IRscores", we select the passage with highest combined score according to BERT+Yes/No and the search engine.

All the reported results have been achieved using the BERT Base Uncased model.

Results Table 4 summarizes our results when training and testing exclusively on DoQA. Note that the performance of all systems in the development set is lower than in the test set, as the latter contains multiple possible answers for the queries (c.f. Section 3.3). Also, for the same reason the F1 column of the development results is empty. Overall the table shows small differences between the BERT and BERT+Yes/No models. However, the BERT+HAE model yields the best results, with an improvement of almost 7 points. This stresses the importance of considering the past history when answering a question in a conversation.

Results of the transfer learning systems are shown in Table 5. The table shows that fine-tuning the model with QuAC alone does not outperform the non-transfer models. However, combining QuAC and DoQA yields to the best results overall, with a gain of almost 6 points w.r.t the non transfer counterpart. This results shows that the information transferred from datasets with different characteristics such as the type of questions is still is beneficial for dealing with domain specific conversational systems.

Model	F1	HEQ-Q	HEQ-D	Y/N	F1-all	F1	HEQ-Q	HEQ-D	Y/N	F1-all
Answer retr.										
Top-1	-	27.2	1.0	78.9	34.7	**41.8**	**38.3**	**6.0**	80.0	**36.6**
Top-20 / BERTprobs	-	25.2	0.5	78.9	30.4	37.1	34.2	4.2	80.0	31.9
Top-20 / BERTprobs & IRscores	-	27.0	0.5	78.9	34.2	40.3	37.2	5.5	80.0	35.0
Question retr.										
Top-1	-	28.8	1.0	78.9	38.5	**45.8**	**41.6**	**7.8**	80.0	**40.9**
Top-20 / BERTprobs	-	26.5	0.5	78.9	33.1	38.9	34.9	4.2	80.0	33.8
Top-20 / BERTprobs & IRscores	-	28.5	0.5	78.9	38.2	45.5	41.1	7.8	80.0	40.5

Table 6: Results (dev on left side, test on right side) of the IR experiments. Here we use the best performing non-contextual model.

Table 6 presents the results of the experiments using the IR rankings. Top-1 approach is the best performing one among the three approaches for both question and answer retrieval strategies. Taking into account the results in Table 1 it was predictable that the question retrieval model was going to outperform the answer retrieval one, however, a greater difference was expected between both of them. This analysis shows that there is a high correlation between the errors of the dialogue system and the retrieval system, making the final performance differences smaller than expected.

8 Conclusion

The goal of this work is to access the large body of domain-specific information in the form of Frequently Asked Question sites via conversational Question Answering (QA) systems. We have presented DoQA (Domain specific FAQs via conversational QA), a dataset for accessing **Do**main specific FAQs via conversational **QA** that contains 1,637 information-seeking dialogues on the cooking domain (7,329 questions in total). These dialogues are created by crowd workers that play the following two roles: the **user** asks questions about a certain cooking topic posted in Stack Exchange, and the **domain expert** who replies to the questions by selecting a short span of text from the long textual reply in the original post. The expert can rephrase the selected span, in order to make it look more natural.

Together with the dataset, we presented results of state-of-the-art models, including transfer learning from Wikipedia QA datasets to our cooking FAQ dataset, and a more realistic scenario where the passage with the answer needs to be retrieved. Our dataset and experiments show that it is possible to access domain specific FAQs with high quality using conversational QA systems with little training data, thanks to transfer learning.

Acknowledgments

This research was partially supported by a Google Faculty Award, ERA-Net CHIST-ERA LIHLITH Project funded by the Agencia Estatal de Investigación (AEI, Spain) project PCIN-2017-118 and the Swiss National Science Foundation (SNF, Switzerland) project 20CH21 174237, the project Deep-Reading (RTI2018-096846-BC21) supported by the Ministry of Science, Innovation and Universities of the Spanish Government, the Basque Government (DL4NLP KK-2019/00045), the UPV/EHU (excellence research group), BigKnowledge - *Ayudas Fundación BBVA a Equipos de Investigación Científica 2018* and the NVIDIA GPU grant program. Jon Ander Campos enjoys a doctoral grant from the Spanish MECD.

References

Chen, D., Fisch, A., Weston, J., and Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., Liang, P., and Zettlemoyer, L. (2018). QuAC: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Dunn, M., Sagun, L., Higgins, M., Güney, V. U., Cirik, V., and Cho, K. (2017). SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine. *CoRR*, abs/1704.05179.

Kočiský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., and Grefenstette, E. (2018). The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv, abs/1611.09268*.

Qu, C., Yang, L., Qiu, M., Croft, W. B., Zhang, Y., and Iyyer, M. (2019). BERT with History Answer Embedding for Conversational Question Answering. *CoRR*, abs/1905.05412.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. *arXiv preprint arXiv:1806.03822*.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Reddy, S., Chen, D., and Manning, C. D. (2018). CoQA: A conversational question answering challenge. *arXiv preprint arXiv:1808.07042*.

Talmor, A. and Berant, J. (2018). The Web as a Knowledge-Base for Answering Complex Questions. 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 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K. (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.

Watanabe, Y., Dhingra, B., and Salakhutdinov, R. (2017). Question Answering from Unstructured Text by Retrieval and Comprehension. *CoRR*, abs/1703.08885.

Yatskar, M. (2018). A qualitative comparison of CoQA, SQuAD 2.0 and QuAC. *arXiv preprint arXiv:1809.10735*.