FOCUSQA: Open-Domain Question Answering with a Context in Focus

Gianni Barlacchi*, Ivano Lauriola, Alessandro Moschitti
Marco Del Tredici, Xiaoyu Shen, Thuy Vu, Bill Byrne and Adrià de Gispert*
Amazon Alexa AI
{gbarlac,lauivano,amosch}@amazon.com
{mttredic,gyouu,thuyvu,willbyrn,agispert}@amazon.com

Abstract

We introduce *question answering with a context in focus*, a task that simulates a free interaction with a QA system. The user reads on a screen some information about a topic and they can follow-up with questions that can be either related or not to the topic; and the answer can be found in the document containing the screen content or from other pages. We call such information *context*. To study the task, we construct FOCUSQA, a dataset for answer sentence selection (AS2) with 12,165 unique ⟨*question, context*⟩ pairs and a total of 109,940 answers. To build the dataset, we developed a novel methodology that takes existing questions and pairs them with relevant contexts. To show the benefits of this approach, we present a comparative analysis with a set of questions written by humans after reading the *context*, showing that our approach greatly helps in eliciting more realistic ⟨*question, context*⟩ pairs. Finally, we show that the task poses several challenges for incorporating contextual information. In this respect, we introduce strong baselines for answer sentence selection that outperform the precision of state-of-the-art models for AS2 up to 21.3% absolute points.

1 Introduction

As more and more information-seeking activities are moving to visual interfaces, the way of interacting with QA systems is changing. An example is given by screen-based virtual assistants (e.g., Google Assistant, Alexa and Siri), where the interaction with the user can be conditioned by the information on the screen.

To study this modeling, we introduce *question answering with a context in focus*, a task where an information-seeking user interacts with a QA system and, after reading some information shown on the screen, they ask a follow-on question. The information may represent the result of a previous exploratory interaction (e.g., *tell me about Lady Gaga*) or by some input provided by the system (e.g., news feeds and daily contents). We refer to this information as the *context* and, for the scope of this work, we define it as the title and the first paragraph of a web page shown on the screen.

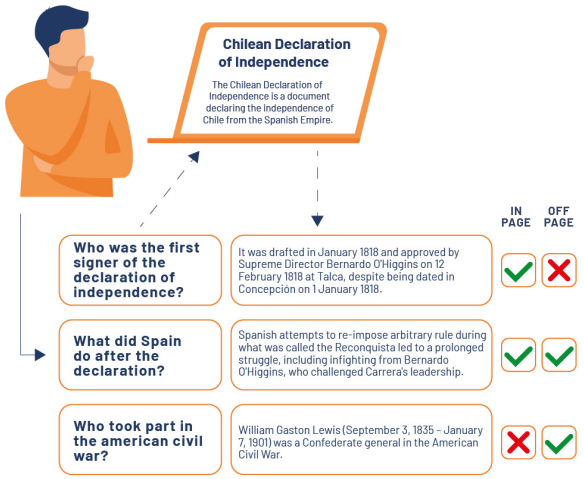


Figure 1: Example of a free information-seeking interaction with a QA system.

Previous works approached this problem as a Machine Reading (MR) task, limiting the application to questions focused on a document (a questioner creates them reading that document) containing the answer (Rajpurkar et al., 2016; Trischler et al., 2017). In contrast, our scenario allows for a free interaction with the system, where users can decide whether to use the context or not. They can ask questions that are (i) *grounded in the context*, e.g., related to entities in the context, (ii) *based on additional knowledge*, i.e., the user has a prior knowledge about the topic and uses it to formulate the question, or (iii) *self-standing*, i.e., related or unrelated to the context, and containing all the information needed to identify the correct answer. All these questions can be answered with in-page text, i.e., from the document, and/or off-page text,

*Corresponding Authors.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5224 - 5237
December 7-11, 2022 ©2022 Association for Computational Linguistics

i.e., from other documents. Figure 1 shows an example of this free information-seeking interaction.

The task poses two key challenges: (i) understanding when and if the question is related to the *context*, and (ii) ensuring that the answers are contextually-relevant when required. In this scenario, ambiguity arises frequently and the answers to the same question may change depending on the context in which they are asked. Even though MR is a common approach to answer extraction for Open-Domain Question Answering (ODQA), we model this problem as an Answer Sentence Selection (AS2) task, where the answer is selected by ranking over all possible candidates (Garg et al., 2020). We believe that AS2 is more relevant to a production scenario since humans converse in compact and complete sentences.

Current QA datasets have limitations that do not allow us to study question answering with a context in focus. They either (i) do not rely on the context at all (Kwiatkowski et al., 2019; Clark et al., 2019; Nguyen et al., 2016), (ii) prompt users to write probing questions based on a given text (Rajpurkar et al., 2016; Trischler et al., 2017), or (iii) assume a conversational user engagement at each turn (Adlakha et al., 2022; Anantha et al., 2021; Choi et al., 2018). To overcome such limitations, we construct FOCUSQA.[1] Our dataset contains 12,165 unique ⟨*question, context*⟩ pairs and a total of 109,940 annotated answers. Instead of asking annotators to write questions, we developed a new methodology that takes existing ones and pairs them automatically with multiple contexts. We sampled questions from NATURALQUESTIONS (Kwiatkowski et al., 2019) (NQ), QReCC (Anantha et al., 2021), and Answer-Sentence Natural Questions (ASNQ) (Garg et al., 2020), a dataset for answer sentence selection derived from NQ.

To show the benefits of our dataset construction methodology, we compare its questions against a sample of 100 ⟨*question, context*⟩ pairs, where the questions are written by humans after reading the *context*. The analysis shows that this common way of crowdsourcing questions leads to examples where all the questions are only related to entities in the context. In contrast, with our approach, we obtain a mix of questions that are grounded in the *context* and questions that contain additional knowledge with respect to the context.

To assess the initial performance on our task/data, we experimented with models for AS2. We show that the task sets several challenges: (i) when models do not use contextual information, state-of-the-art systems only achieve precision of 29.37%. (ii) When we introduce a set of strong baselines that incorporate the *context* and additional context from the answer page into Transformer models (Vaswani et al., 2017), the combination of these contexts achieves a precision of 50.7%. However, the task requires further research exploration to get closer to humans' precision of 81.4% on the test set.

In brief, our contribution is threefold:

- We study a realistic scenario where users can freely interact with a QA system that provides contextual information to facilitate information seeking.
- We construct FOCUSQA, a dataset for AS2 with 12,165 unique ⟨*question, context*⟩ pairs and 109,940 annotated answers. To build the dataset, we developed a novel methodology to elicit more realistic questions.
- We introduce strong baselines that incorporate context from different sources (e.g., screen content and answer page) into Transformer models and show that contextual sentence selection models outperform the precision of state-of-the-art models for AS2 up to 21.36% absolute points.

2 Related Work

Contextual Question Answering (CQA). Contextual Question Answering (CQA) leverages additional context instead of treating the questions as self-contained inputs. The context consists of many factors such as cognitive and social ones that are related to a user's intentions, tasks, and needs (Allen, 1997). Taking the context into account is crucial to better interpret the questions (Min et al., 2020). Earlier works and datasets focused on leveraging external context like the time, location, and user profiles (Krulwich and Burkey, 1997; Limbu et al., 2009; Zhang and Choi, 2021). Recently, the advance of various smart devices allows users to interact with the system in a much richer way. This has influenced the creation of CQA datasets leveraging a wider forms of context such as newspapers (Trischler et al., 2017), web pages (Chen et al., 2021), visual information (Zhu

[1] https://github.com/alexa/focusqa

et al., 2016; Biten et al., 2019), and conversational history (Choi et al., 2018; Qu et al., 2020; Anantha et al., 2021). Trischler et al. (2017) is the closest to our research in that we use the same form of context. However, they collect questions by showing context to annotators explicitly, which leads to shallow, superficial questions far from real scenarios.

Contextual Sentence Selection (CS2). The AS2 task was originally defined in the TREC competition (Wang et al., 2007) and has the advantage of high efficiency, which enables its use in real-world applications (Garg et al., 2020). Previous AS2 models treated each sentence as an independent unit and applied neural network models to select the sentence with the highest score (He and Lin, 2016; Yang et al., 2016; Garg et al., 2020). However, such approaches are sub-optimal as sentences extracted from web documents are typically not self-contained and they can have issues related to pronouns, anaphora, or other linguistic limitations (Tan et al., 2017). Contextual Sentence Selection (CS2) addresses these limitations by encoding additional information, i.e., context, into the scoring function. Recently, Lauriola and Moschitti (2021) showed that local context, defined as the two sentences surrounding the candidate answer, highly improves non-contextual state-of-the-art models. Han et al. (2021) proposed to use non-consecutive multi-sentences coming from the document containing the answer. Differently from these approaches, this paper defines, analyzes, and evaluates cross-document context by combining the context in focus and retrieved documents simultaneously.

3 FOCUSQA Setup

3.1 Defining Context in Focus

In this work, we consider the screen content as a context. We say that a question is *context-dependent* if its answer a_i can change depending on the context c_i. Typically, such questions contain pronouns, entities, and concepts that can be disambiguated only if the context is provided. In contrary, we say that a question is *context-independent* if it provides all the information needed to find an unambiguous answer. In this work, the context is defined as the *title* and the *first-paragraph* of a web page. See Table 1 for examples in which the same questions can have different answers depending on the context.

3.2 FOCUSQA task: QA with a Context in Focus

The task simulates the real use-case scenario where there is a user and a context in focus. The user is free to interact with the system and after reading the context, they can ask a question that is either *context-dependent* or not. The answer can be extracted from the document in focus or from other pages of the web. We call *in-page* candidates those extracted from the document containing the context in focus and *off-page* candidates those extracted from other pages. We cast this problem as an AS2 task and we extend the formulation to include the context in focus. Thus, given a question q and a set of answer candidates $\mathcal{S} = \{s_1, ..., s_n\}$, the task of *QA with a context in focus* is to select a sentence s_i that correctly answers q for the provided context c.

Formally, let $\mathcal{Q}$ be the set of questions, $\mathcal{C}$ the set of contexts paired with the questions, and $\mathcal{S}$ the set sentences. The task can be defined as a ranking problem. Given a pair $(q, c) \in \mathcal{Q} \times \mathcal{C}$ and a set of possible answers $\mathcal{S}_{(q,c)} \subseteq \mathcal{S}$ for q given c, the sentence selector returns the answer $a \in \mathcal{S}_{(q,c)}$ for which

$$a = \arg \max_{s \in \mathcal{S}_{(q,c)}} r(q, s, c)$$

where r is the scoring function $r : \mathcal{Q} \times \mathcal{S} \times \mathcal{C} \rightarrow \mathbb{R}$, which can be estimated using Transformer models, as explained in Section 6.

4 Data Collection

Our approach to build the FOCUSQA dataset is divided into two stages: (i) *question/context pairing*, and (ii) *answer collection*. To collect candidate answers we implemented an Open-Domain Question Answering (ODQA) system for AS2. For this work we used a BM25/lexical-based sentence retriever and a Tranformer-based AS2 reader, as this combination already implements a reliable QA system. We retrieve documents using an open-domain index containing ∼100M web pages from the open repository of Common Crawl.[2] See Appendix A for more implementation details. All the annotation tasks are performed by a team of professional annotators with a project lead[3]. Additional details on the guidelines can be found in Appendix A.

[2] https://commoncrawl.org/

[3] The team is part of a company that offers professional data labeling services.

Question	Context (*title*, first paragraph)	Answer
Who is considered the author of the constitution?	*The Scheduled Tribes In India.* The Scheduled tribes are groups of people that are officially recognized in the Indian Constitution. They are also referred to as Dalit, which translates as broken or scattered.	Ambedkar, who lived from 1891 to 1956, was an Indian economist and is considered the father of the modern Indian Constitution.
	Constitution of United States of America 1789. While the primary authorship of a vast array of documentations and publication may be cited with ease, the process of identifying the Father of the Constitution may prove to be a far more difficult ...	James Madison – alongside fellow Federalist Alexander Hamilton – is considered to be one of the individuals credited with being the Father of the Constitution.

Table 1: Example of a question with answers that should be chosen depending on the context.

4.1 Question/Context Pairing.

We construct ⟨*question, context*⟩ pairs by sourcing questions from NATURALQUESTIONS (Kwiatkowski et al., 2019), QReCC (Anantha et al., 2021), and Answer-Sentence Natural Questions (ASNQ) (Garg et al., 2020).

Instead of requesting annotators to formulate questions about the context, we propose to link the question to a valid context by working backwards from candidate answers. For each question, we use the end-to-end ODQA system to collect a ranked list of candidate sentences. Then, given a question q and an answer sentence from the top-k scored candidates s_i, with $i = 1, ..., k$, we pair q with the *title* and the *first-paragraph* of the web page that contains s_i. We pair each question with up to 3 different contexts using the top-3 answers extracted from different documents. Because this is an automatic procedure, we ask annotators to validate each ⟨*question, context*⟩ pair with three annotations:

- *context-dependent*: the answer to the question can change depending on the context.
- *context-connected*: the question refers to concepts, entities, or events related to the context. Most likely, the answer to the question can be found in-page or in documents with similar contents.
- *context-answered*: the question is partially or fully answered by the context.

We consider valid a ⟨*question, context*⟩ pair if it is *context-dependent* or *context-connected*, but not *context-answered*. We stop the annotation if any of these conditions are not satisfied.

In addition, we collected self-standing questions from the test set of ASNQ. Given the question and the context from its answer page, we compute the similarity of such context with a set of contexts retrieved from the index. Then, we pair the question with the least and the most similar contexts. This way, the resulting pairs contain questions that can be answered without additional context, posing the challenge for the model to recognize when the context is needed or not.

Split	Q	Q/C	Q/C/A	Neg.	Pos.	Avg. Candidates
All	6,036	12,165	109,940	92,264	13,643	9.0
Train	3,276	6,756	49,386	39,867	7,278	7.3
Dev	800	1,698	12,498	10,208	1,781	7.4
Test	1,960	3,711	48,056	42,189	4,584	12.9
Test (contextual)	1,250	2,711	27,264	23,063	2,918	10.1
Test (self-standing)	710	1000	20,792	19,126	1,666	20.8

Table 2: Data Statistics.

4.2 Answers Collection.

After identifying valid ⟨*question, context*⟩ pairs, we collect multiple triplets ⟨*question, context, answer*⟩ and annotate them via crowdsourcing.

Annotators are asked to judge if the candidate sentence (i) is about the same context of the question, (ii) answers the question, and (iii) is factually correct. We consider an answer correct if all conditions are true. We collect sentences from two sources:

- *in-page*, where we select the top-k candidates from the document in focus.
- *off-page*, where given the question and the context, i.e., *title* and *first-paragraph*, we use the ODQA system to collect candidate answers. We perform a basic contextual retrieval by querying the index using the concatenation of the question and the title.

For the set of self-standing questions, we collected candidates in the same way. In addition, we used candidates from the answer page. Instead of re-annotating the pairs, we rely on the annotation from ASNQ and we labelled as negative all the other pairs. Considering that these questions are not context-dependent, we can assume that retrieved candidates are most likely coming from unrelated documents. More details can be found in Appendix A.3.

	in-page		off-page		both	
	#	%	#	%	#	%
All	6,546	53.8	1,321	10.8	966	7.9
Test	1,619	43.6	1,088	29.3	217	5.8

Table 3: Sources of answerable ⟨*question, context*⟩ pairs. The percentage is computed respect to the total number of answerable pairs.

Question	Context (*title*, first paragraph)
P: What age did he join the *Premier League*? **M:** Has Delial Brewster played for the England national team?	*Delial Brewster* Delial Brewster (born 7 November 1997) is an English professional footballer who plays as a forward.
P: Was *Carson* injured in the attack? **M:** Who directed Dirty Pair: Project Eden?	*Dirty Pair: Project Eden* Meanwhile, on the planet of Agerna, one of the planet's two major refinery factions is the subject of a vicious attack, and is taking some heavy damage.

Table 4: Questions with expertise: **P** are our questions; in underline the new entities; **M** are questions by human annotators.

5 Data Analysis

5.1 Dataset Statistics

The final dataset contains 6,036 questions, which we paired with 12,453 context in focus. We obtained 12,165 unique ⟨*question, context*⟩ pairs, for which we collected a total of 109,940 answers. Among these, 89,148 are new human annotations, 6,666 are from ASNQ, and 14,126 are automatically labelled as negative. We split the data into train/dev/test by randomly sampling unique questions from the set, excluding all the self-standing questions. We used them only at test time to evaluate the model's ability to distinguish when the context is not necessary. We refer to this subset as self-standing, i.e., questions that do not require the context to be answers, in contrast to contextual. Table 2 shows the statistics of our collected dataset.

The dataset contains 81.4% of answerable ⟨*question, context*⟩ pairs. Among these, in Table 3 we can observe that both in-page and off-page are valuable sources to find the correct answers. This is reasonable as questions are asked in an information-seeking scenario and users can ask follow-on questions based on the context. Finally, it is worth mentioning that our approach for automatically creating ⟨*question, context*⟩ pairs allows for saving annotation cost up to 31%. More details can be found in Appendix A.6.

5.2 Questions Based on Additional Knowledge

By design, our dataset, introduces an interesting feature that elicits more realistic questions, namely, some of the questions imply *topic knowledge* from the questioner. We refer to them as *questions based on additional knowledge*. When collecting questions manually, it is very complicated to obtain these questions since annotators might not have specific expertise about the topic. To demonstrate this, we randomly sampled 100 contexts from our dataset. For each of them, we asked annotators to formulate natural and well-formed questions that are not answered with the context. Consider the example in Table 4 where (P) is the automatically paired question and (M) is the one asked looking at the context. (P) is a question that can be only asked if one knows when "Brewster joined Premier League", because that is not specified in the context. Conversely, (M) are firmly grounded in the context and do not reveal any background knowledge, besides general world knowledge (e.g., England has a national team).

To measure the topic expertise of paired and manual questions, we compute the percentage of entities in the questions that are not mentioned in the *context*. As shown in the example in Table 4, mentioning a new entity is a reliable proxy of domain specific knowledge. We observe that only 6% of the entities in (M) questions are not mentioned in the context, while for (P) it goes up to 59%. Additionally, we compute the semantic similarity between questions and context vectors, which we obtained using Sentence-BERT, a state-of-the-art model for sentence embedding (Reimers and Gurevych, 2019). The average similarity between (M)/(P) questions and their context is 0.51/0.37. This result confirms that our dataset includes questions whose questioner owns more knowledge on the topic than what can found in the context.

5.3 Differences with Other QA Datasets

In this section we discuss the difference between FOCUSQA and Machine Reading Comprehension (MRC) datasets, where data are also in the form of ⟨*question, context*⟩ pairs. First, FOCUSQA is intended to simulate users interactions with a screen-based QA system, which is a real-world application not modeled by current MRC datasets. The latter are created by providing a text to the annotators and asking them to generate questions. As a consequence, and as discussed above, the generated questions are strictly grounded in the text, i.e., related to text entities, and so these questions have a high lexical overlap and similarity with the text (context).

Question	Context (*title*, first paragraph)	In-Page	Off-Page
What records did he produce?	*Tony Berg.* Anthony Rains "Tony" Berg (born October 21, 1954 in Connecticut) is an American musician, record producer, and A&R representative, in which role he has been described as an "industry guru".	✓	✗
Were there many casualties?	*What Countries Fought in World War II? | Reference.com.* The countries that fought in World War II were Germany, Italy and Japan, which comprised the Axis Powers, and Britain, France, Australia, Canada, New Zealand, India, the Soviet Union, China and the United States of America, which comprised the Allies.	✗	✓
When will the capsule be opened?	*List of time capsules.* This is a list of time capsules. The register of The International Time Capsule Society estimates there are between 10,000 and 15,000 time capsules worldwide. An active list of Time Capsules is maintained by the Not Forgotten Digital Preservation Library.	✓	✓
How was Santana experimenting?	*Fredo Santana.* Derrick Coleman (July 4, 1990 – January 19, 2018), known professionally as Fredo Santana, was an American rapper from Chicago, Illinois.	✗	✗

Table 5: Example of answerable ⟨*question, context*⟩ with answers found in different sources.

To obtain more realistic data, FOCUSQA takes a novel approach where questions are linked to a valid text (i.e., the context) that contains the correct answer to the questions. The approach is general and it can be applied to any collection of questions, avoiding the problem of having only questions specific to the context. Second, FOCUSQA contains correct answers that are extracted from multiple documents (Table 5), whereas in MRC datasets (e.g., NewsQA (Trischler et al., 2017), SQuAD (Rajpurkar et al., 2016) and NATURALQUESTIONS (Kwiatkowski et al., 2019)), answers are extracted only from the documents used to generate the question. Finally, in MRC-style datasets (e.g., SQuAD and NewsQA), the referring text is closely related to the question and it might even contain the answer (e.g., SQuAD) or it is needed to answer the question (e.g., NewsQA). By contrast, FOCUSQA models the situation where a user reads a context and asks questions that may or may not be related to the context (topic switching). This is essential to model user behavior when interacting with screen-based QA systems. To be successful in FOCUSQA, QA systems must learn when and how to exploit context in focus.

6 Contextual Models

Contextual Sentence Selection (CS22) models are Transformer models with multiple token-type (sentence) embeddings (Lauriola and Moschitti, 2021) and as in input a single textual sequence of question, answer, and context. They are very suitable to be used for tasks where the context is required to find the answer. To adapt such models to the context available in FOCUSQA, we introduce two new input sequences, namely, *READOUT* and *Cross-Context*, which embed question/answer pairs along with text from the answer page and the textual *context* from the screen, respectively.

REAd DOcument Until Truncation (READOUT). We hypothesize that document-level textual information can orthogonally help the sentence selector. The *READOUT* context encodes the document until truncation, i.e., *[CLS] question [SEP] answer [SEP] document [EOS]*, depending on the maximum sequence length:

- $READOUT_Q$, where the model encodes the document in focus.
- $READOUT_A$, where the model encodes the document from which the candidate answer sentence is extracted and ignore the context of the question.

Cross-Context. Documents can be really long and their ingestion into a Transformer model can dramatically increase the latency in both training and inference. To overcome these limitations, we propose to model both question and answer context, i.e., *title* and *first paragraph*, from the screen and from the answer page, using more compact text.

- $Cross\text{-}Context_{Titles}$, where titles from the *context* and the answer page are used. The information is encoded into a Transformer model as *[CLS] question [SEP] answer [SEP] context title [SEP] answer document title [EOS].*
- $Cross\text{-}Context_{QA}$, where the model combines the context with the answer document in a single model. An example is encoded as *[CLS] question [SEP] answer [SEP] title [SEP] first paragraph [SEP] answer title [SEP] answer document first paragraph [EOS].*

Note that *Cross-Context* models exploit extra information coming from both answer and question documents (*cross*-documents), whereas *READOUT* uses a single document depending on the specialization.

Model	P@1	MRR	HIT@3
ORACLE	83.0	-	-
No-Context	$29.37_{\pm 0.6}$	$44.25_{\pm 0.5}$	$52.19_{\pm 0.3}$
Local	$33.56_{\pm 0.6}$	$47.42_{\pm 0.4}$	$55.62_{\pm 0.0}$
READOUT$_{\text{A}}$	$25.77_{\pm 0.5}$	$40.64_{\pm 0.4}$	$48.61_{\pm 0.3}$
READOUT$_{\text{Q}}$	$26.58_{\pm 0.3}$	$40.78_{\pm 0.3}$	$47.05_{\pm 0.6}$
Cross-Context$_{\text{Titles}}$	$\mathbf{50.73_{\pm 0.7}}$	$\mathbf{59.36_{\pm 0.7}}$	$\mathbf{64.15_{\pm 1.0}}$
Cross-Context$_{\text{QA}}$	$50.03_{\pm 2.8}$	$57.93_{\pm 3.0}$	$61.38_{\pm 4.8}$
No-Context$_{\text{QR}}$	$33.50_{\pm 0.8}$	$47.72_{\pm 0.6}$	$56.07_{\pm 0.4}$
Cross-Context$_{\text{QR+Titles}}$	$47.70_{\pm 0.8}$	$56.49_{\pm 0.7}$	$63.45_{\pm 1.1}$

Table 6: P@1 computed by contextual models and baselines on FOCUSQA. Best results are highlighted in bold.

Model	in-page	out-page
No-Context	28.08	49.91
Local	30.99	58.52
READOUT$_{\text{A}}$	29.22	35.02
READOUT$_{\text{Q}}$	42.04	17.31
Cross-Context$_{\text{Titles}}$	89.56	22.21
Cross-Context$_{\text{QA}}$	90.01	19.12
No-Context$_{\text{QR}}$	37.41	48.74
Cross-Context$_{\text{QR+Titles}}$	83.56	22.21

Table 7: P@1 on questions answerable only with in-page candidates or with off-page candidates.

7 Experiments

7.1 Experimental Setup

Baselines. We use state-of-the-art models for AS2 (Garg et al., 2020) and Contextual AS2 (Lauriola and Moschitti, 2021). We refer to this model as *No-Context* and *Local*.

Question Rewriting. In order to include contextual information in AS2 models, we performed Question Rewriting (QR) to reformulate the questions to a more self-contained form. The questions are obtained with a generative model for question rewriting (Anantha et al., 2021) using the concatenation of the question with the *context*.

Transfer and Adapt with Context. We trained our models adopting a *Transfer and Adapt* strategy (Garg et al., 2020): first (Transfer), we fine-tune a pre-trained ELECTRA-base[4] model on the large ASNQ dataset (20M q/a pairs). Then (Adapt), we further fine-tune on FOCUSQA. More details can be found in B.

7.2 Results

Results of the proposed models and baselines are reported in Table 6. We use standard evaluation metrics for AS2: Precision@1 (P@1), Mean Reciprocal Recall (MRR) and hit rate at k (HIT@3). We can observe that models that do not mix question context and answer context perform the worst. This result is expected as the answer can be context-dependent. We note that: (i) READOUT$_{\text{A}}$ and READOUT$_{\text{Q}}$ have the lowest performance with a P@1 of 25.77% and 26.58%, respectively. This is expected as they heavily rely only on one of the two contexts; (ii) when no context is used at all, performances are better (29.37%) and they go up to 33.56% with *Local*, thanks to the information coming from the text surrounding the answer; (iii) the highest performances are achieved by Cross-Context$_{\text{Titles}}$ (50.73%) and Cross-Context$_{\text{QA}}$ (50.03%) models, which exploit information from both question and answer contexts.

In the last two rows of Table 6, we used the rewritten questions instead of the original ones. This enables the standard AS2 model to use the context and to combine question rewriting and CS2 models to evaluate if having self-contained questions helps. We observe that using rewritten questions improves by 4.13% when no context is provided. In contrast, in line with results in Del Tredici et al. (2021), we found that question rewriting is not useful if the context is already taken into account by other means.

To investigate more on the behavior of contextual models, in Table 7, we split the dataset considering questions answerable with candidates from in-page and off-page only, respectively. *No-Context* and models with only the answer context have low performances when the answer is in-page. This is because in-page candidates are extracted from the document in focus. In contrast, off-page candidates are retrieved by a lexical-based search engine, and they tend to have more overlap with the question. Thus, for the model, it is more challenging to contextualize in-page answers and rank them to the top when they are correct. We can also observe that the performance of *Cross-Context* models are very optimized to answer questions with in-page candidates. However, they struggle when context is different and, possibly, unrelated. This illustrates the challenges set by the task: (i) understanding when a question is related to the context and (ii) selecting answers that are contextually-relevant.

[4] available from HuggingFace.

Question	Context (*title*, first paragraph)
Q: Where is the start of 17 mile drive? **QR**: Where is the start of 17 mile drive from emery county, utah?	*Rochester Rock Art Panel* The Rochester Rock Art Panel in Emery County, Utah consists of a large number of petroglyphs of various ages. Some are prehistoric rock art, probably of Fremont culture origin. Others are probably modern, depicting horses, for example.
Q:Which of his writing was mentioned in the page? **QR**:Which of Sir William Petty's writing was mentioned in the Declaration Concerning the newly invented Art of Double Writing page? **Q**: What kind of book is where's waldo **QR**: What kind of book is where's waldo at circus	*Double Writing (Petty).* A Declaration Concerning the newly invented Art of Double Writing was a pamphlet of 6 leaves, written by Sir William Petty (1623-1687) and first published in 1648. It contained information regarding his invention of the "Art of Double Writing". *Where's Waldo at the Circus* Designed for "children ages 4 through 8", Where's Waldo at the Circus is a computer video game that immerses the player in a rich interactive environment complete with music, sound, and animation.
Q: a type of basic rock popular in the 1980s **QR**: This is a list of rock music genres consisting of subgenres of popular music that have roots in 1940s and 1950s rock and roll, and which developed into a distinct identity as rock music in the 1960s, particularly in the 1980s	*List of rock genres* This is a list of rock music genres consisting of subgenres of popular music that have roots in 1940s and 1950s rock and roll, and which developed into a distinct identity as rock music in the 1960s.

Table 8: Examples of errors when using the context to rewrite questions.

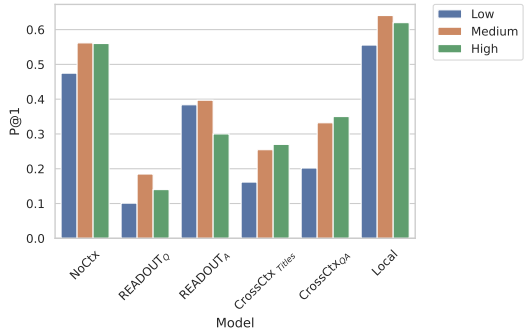


Figure 2: Models performances on self-standing depending on the similarity between question and context.

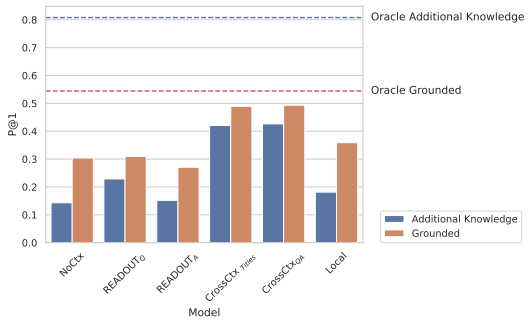


Figure 3: Models performances on a set of grounded questions and questions based on additional knowledge.

8 Error Analysis

In this section we further analyse models behavior depending on the type of question and context.

Do models performance vary when context is not needed? Results from Table 7 highlight that models have different performances depending on the source of the answer. However, it is unclear how the performances vary with self-standing questions that do not require the context to be answered. To further investigate this, we focus on the self-standing questions in our test set. Given a question, we measure the similarity with its context, using the sentence embedding obtained with Sentence-BERT (Reimers and Gurevych, 2019). Then, based on the similarity score, we split the 1,000 ⟨*question, context*⟩ pairs into three groups based on the percentile: *low*, which contains the lowest 10%; *high*, which contains the highest 10%, and *medium* with the remaining pairs.

In Figure 2, not surprisingly, we can observe that models that do not rely on the context, i.e., *No-Context* and *Local*, are more robust to these questions. Instead, all the other models perform considerably worst. In particular, *READOUT$_Q$* has the lowest performance as it heavily relies on the question context. Finally, we observe that all the models have a lower precision when the question is less related to the context. These findings suggest that CS2 models have a subpar performance on questions that do not require the context, i.e., self-standing questions, and a different modeling is required to make the best usage of the context.

How do models perform on questions based on additional knowledge? While having questions based on additional knowledge is more realistic, it is unclear what new challenges they introduce in the task. As described in Section 5.2, measuring the percentage of new entities in the questions compared to the context is a reliable proxy to spot questions based on additional knowledge. We considered all the questions in our test set that have more than one new entity, obtaining a set of 211 ⟨*question, context*⟩ pairs. Then, from the remaining set, we randomly sampled 250 *grounded questions*. Figure 3 shows the models performance on these two sets of questions. We note that: (i) questions based on additional knowledge are more challenging for the retrieval, with 54% of them that are an-

swerable, versus the 81% of answerable grounded questions; (ii) models have more difficulty in finding the correct answers, with a gap in P@1, respect to grounded questions, that goes from 8% with Cross-Context$_{\text{Titles}}$ to 16% when no context is used. Questions based on additional knowledge might be asked during information-seeking interactions, with results showing that for state-of-the-art models such questions still represent a challenge. This is a relevant finding from a practical point of view, since it provides a valuable indication to design QA datasets that better capture realistic interactions.

How does Question Rewriting perform when using the context? In Del Tredici et al. (2022), authors show that QR can fail when a large amount of information is required. From a manual inspection, we found this problem to be present also in QA with a context in focus. Table 8 shows some of these examples. We can note that, not only QR fails by providing very long and convoluted rewrites (row 3), but it also uses the context when not useful. For example, in row 2 and 3 the question is self-contained, i.e., *17-Mile Drive is a scenic road in California, US*, and adding *emery county, utah* invalidates the question. While QR helps AS2 models, our results seems to indicate that understanding when to rely on the context is remains a crucial challenge in QA with a context in focus.

9 Conclusion

We introduced *question answering with context in focus*, a task where there is a context in focus and users can ask questions that can be answered from in-page sentences, i.e., from the document in focus, or from off-page sentences, i.e., from other documents. To study the task, we constructed FOCUSQA, a dataset with 12,165 unique ⟨*question, context*⟩ pairs and a total of 109,940 answers. In order to elicit more realistic questions, i.e., annotators ask a question only based on that text information, we proposed a new methodology that can take any existing question and automatically pair it with a context in focus. We also introduced new input sequences for CS2 models. Our experiments show their effectiveness in learning from our data, as they greatly outperform state-of-the-art models for AS2 that do not make use of context. FOCUSQA highlights challenges of modeling realistic information-seeking scenarios and invites further research into this area. For example the retrieval struggles to extract contextually-relevant passages from an open-domain index. Future research can be devoted to study new approaches to model the context in focus in the retrieval stage. While we studied how to inject cross-contexts into QA models, learning when to use or not the context is another future research direction to explore. Furthermore, context is limited to text (i.e., title and first paragraph) and future research may include extending the task to a multi-modal or multi-turn scenario.

10 Limitations

We acknowledge this work to be limited in three aspects. First, some of the candidates from ASNQ have been automatically labelled as negative. While we did apply several measures to mitigate this problem, there exists the possibility of having some false-negative among the candidates. Second, while the retrieval is not the core of this work, we did implement it as a part of the ODQA system that collects candidate sentences. However, our basic implementation of the retrieval struggles to extract contextually-relevant passages. As a result, in off-page candidates there is a high percentage of negative samples. We opted for a sparse retrieval, i.e., BM25, because it is a common approach, robust on noisy web data, that can be used to more efficiently index a large set of domain (e.g., Common Crawl). Nonetheless, using dense models (Shen et al., 2022) is an important future work that we plan to explore to improve the efficiency of our methodology to build contextual datasets. Finally, even though our index greatly supports the scope of this work, a more variety of websites (e.g., news websites) can help collect less entity-centric contexts.

11 Ethics statement

This work relies on the publicly available datasets for Open-Domain Question Answering. The dataset proposed in this work can be helpful in advancing the research in QA and Conversational QA. Even with our best efforts to ensure the quality of the content, answers extracted from webpage may have a biased view, for example political opinions. The work does not propose models that can generate harmful or toxic content. Our models are fine-tuned based on checkpoints downloaded from HuggingFace (Wolf et al., 2020).

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. Topiocqa: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Bryce Allen. 1997. Information needs: A person-in-situation approach. In *Proceedings of an international conference on Information seeking in context*, pages 111–122.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

Xingyu Chen, Zihan Zhao, Lu Chen, JiaBao Ji, Danyang Zhang, Ao Luo, Yuxuan Xiong, and Kai Yu. 2021. Websrc: A dataset for web-based structural reading comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4173–4185.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau 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.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936.

Marco Del Tredici, Gianni Barlacchi, Xiaoyu Shen, Weiwei Cheng, and Adrià de Gispert. 2021. Question rewriting for open-domain conversational qa: Best practices and limitations. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2974–2978.

Marco Del Tredici, Xiaoyu Shen, Gianni Barlacchi, Bill Byrne, and Adrià de Gispert. 2022. From rewriting to remembering: Common ground for conversational qa models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 70–76, Dublin, Ireland.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7780–7788.

Rujun Han, Luca Soldaini, and Alessandro Moschitti. 2021. Modeling context in answer sentence selection systems on a latency budget. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3005–3010, Online. Association for Computational Linguistics.

Hua He and Jimmy Lin. 2016. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *Proceedings of the 2016 conference of the north American chapter of the Association for Computational Linguistics: human language technologies*, pages 937–948.

Bruce Krulwich and Chad Burkey. 1997. The infofinder agent: Learning user interests through heuristic phrase extraction. *IEEE Expert*, 12(5):22–27.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Ivano Lauriola and Alessandro Moschitti. 2021. Answer sentence selection using local and global context in transformer models. In *European Conference on Information Retrieval*, pages 298–312. Springer.

Dilip K Limbu, Andrew M Connor, Russel Pears, and Stephen G MacDonell. 2009. Improving web search using contextual retrieval. In *2009 Sixth International Conference on Information Technology: New Generations*, pages 1329–1334. IEEE.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPs*.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 539–548.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Xiaoyu Shen, Svitlana Vakulenko, Marco del Tredici, Gianni Barlacchi, Bill Byrne, and Adrià de Gispert. 2022. Low-resource dense retrieval for open-domain question answering: A comprehensive survey. *arXiv preprint arXiv:2208.03197*.

Chuanqi Tan, Furu Wei, Qingyu Zhou, Nan Yang, Bowen Du, Weifeng Lv, and Ming Zhou. 2017. Context-aware answer sentence selection with hierarchical gated recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(3):540–549.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Mengqiu Wang, Noah A Smith, and Teruko Mitamura. 2007. What is the jeopardy model? a quasi-synchronous grammar for qa. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 22–32.

T Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, et al. 2020. Huggingface's transformers: State-of-the-art natural language processing. arxiv 2019. *arXiv preprint arXiv:1910.03771*.

Liu Yang, Qingyao Ai, Jiafeng Guo, and W Bruce Croft. 2016. anmm: Ranking short answer texts with attention-based neural matching model. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 287–296.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. *arXiv preprint arXiv:1902.01718*.

Michael JQ Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa. *arXiv preprint arXiv:2109.06157*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

A Data Collection Details

A.1 ODQA System for AS2

We implemented a standard ODQA system for AS2 that computes answers in three phases: (i) text retrieval, which returns relevant documents for a question from a large text collection; (ii) text ranking, which reranks and decomposes text into answer candidates, e.g., sentences; and (iii) answer sentence selection (QA), which selects the final answer for a question from the list of candidates.

On the same line of similar end-to-end systems for QA (Yang et al., 2019), we implemented text retrieval with a BM25 ranking function. To split the text into sentences we used an off-the-shelf sentence splitter (Manning et al., 2014). Sentences are ranked using a state-of-the-art Transformer-based model for AS2 (Garg et al., 2020). Text retrieval is done on a standard index using Lucene/Elasticsearch. As a framework for text ranking we used HuggingFace (Wolf et al., 2020). The tokenizer is set to truncate to a maximum length of 128 tokens, removing a token from the longest sequence in the input.

A.2 Document Collection

The index is built using a large collection of Web data, i.e., documents. This resource allows us to measure the impact of our work in an industry-scale ODQA setting. We selected English Web documents of the 5,000 most popular domains, including Wikipedia, from releases of Common Crawl in 2019 and 2020. This process produced a collection of $\sim$100M of documents. Each document in the index contains a url, title of the page, and content of the document, after removing all the HTML tags.

A.3 Pre-processing and Filtering

When collecting answer candidates, to optimize the annotation cost, we rely on the model score

and we send for annotation only sentences with $score > 0.25$. We discarded questions where: (i) we have less than 3 contexts, i.e., we do not have at least three answers with $score \leq 0.25$ to use for contex pairing; (ii) we have less than 3 candidate answers to annotate from *in-page*; and (iii) we have less than 3 candidate answers to annotate from *cross-page*. To obtain the first paragraph of the page, we split based on a double newline character. Then, we remove the title (if present) from the text and we truncate the text after 40 words. We discarded a context if it has less than 10 words. We always retrieved 100 documents from the index. To collect in-page candidates we used $k = 3$. For cross-page candidates, we use $k = 10$ for training examples, and $k = 15$ for testing examples. To sample question/context pairs from ASNQ, we filter out pairs with less than 10 candidates and more than 25. In retrieval, we limited the index to only Wikipedia pages. In this way, we reduce the possibility to retrieve pages that are similar to the answer page. Because those candidates for ASNQ are automatically labelled as negative, there is still the possibility to have false-negative.

A.4 Dataset Composition

The dataset is built by aggregating publicly available data from existing repositories and QA datasets. Questions are sourced from NATURALQUESTIONS (Kwiatkowski et al., 2019) and QReCC (Anantha et al., 2021). Answers are extracted from web pages contained in the Common Crawl index. To create the subset of *context-independent* ⟨*question, context*⟩ pairs, we sample used data from the ASNQ dataset (Garg et al., 2020). For each entry, we will release the annotation labels, the question and answer document id, and an identifier for the questions to allow the match with the source dataset. We used the MD5 hashing function from the Python package *hashlib*.

A.5 Annotation Guidelines and Tooling

All the annotation tasks conducted in this paper are performed by a team of professional annotators with a project lead. Annotation is performed using a custom annotation interface based on the annotation guidelines (Figure 4).

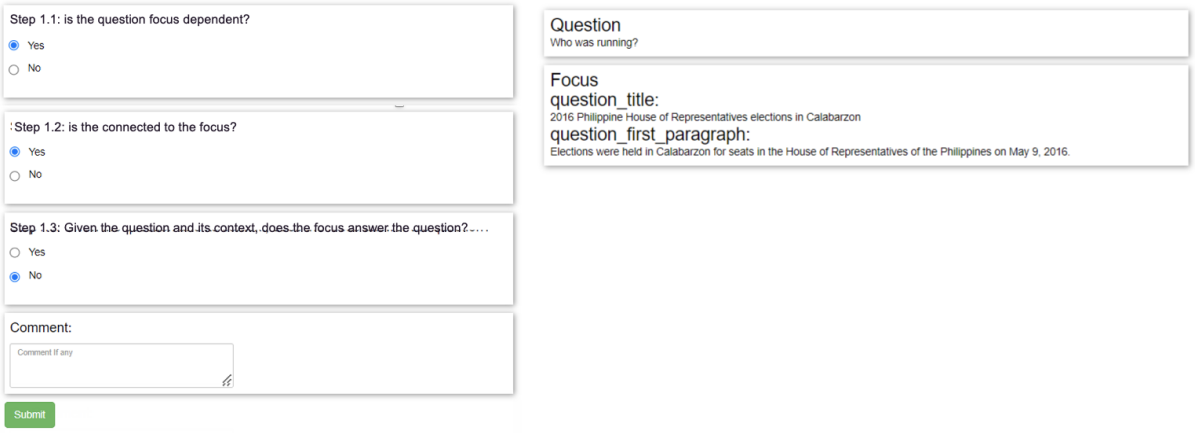
Guidelines. We provided annotators with guidelines to follow during the process. The guidelines describe the two annotation steps for ⟨*question, contex*⟩ validation and answer collection. For each step and substeps, we provided extensive examples, covering edge-cases as much as possible. When evaluating if a question is context-connected, annotators also check if ⟨*question, context* ⟩ presents grammar inconsistencies (e.g., pronouns in the question that are inconsistent with the person of the entities in the context) and if the question is about possible facts. The guidelines were designed with an iterative approach with three pilots. For the pairs in the pilot, two authors of the paper conducted a separate annotation that was used to assess the quality of the pilot. We manually reviewed the annotations and we discussed with the annotation project lead all the problems found in the process. We improved the guidelines according to their feedbacks and we ran the annotation at scale once we obtained an agreement greater than 90% with our manual annotations.

Tooling. The high presence of ambiguity makes the task hard even for humans; it is difficult to know if a candidate sentence can answer the question without considering additional contextual information. For this reason, annotators are provided with the context and additional text from the document containing the answer, i.e., title and the paragraph containing the answer (Figure 4). Annotators are trained on the specific task prior to start the annotation task. The process is divided into two conceptual stages: (i) question/context pairing, and (ii) answer collection. First, annotators have to validate a ⟨*question, context*⟩ pairs. Then, only if the pair is valid, they are asked to annotate the candidate answers. At each moment in the annotation process, annotators are allowed to use a commercial search engine to clarify the content of questions, context, and answer.

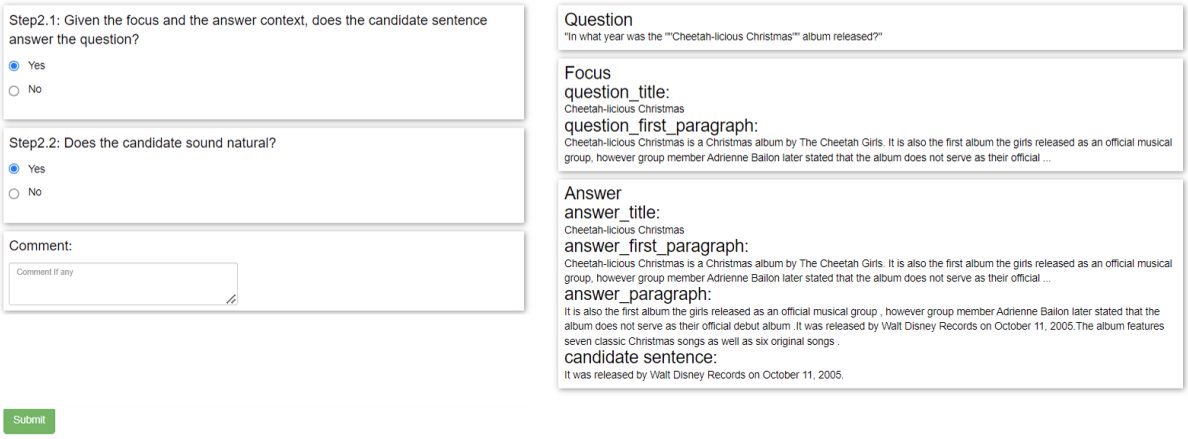
Quality Control To ensure high quality of annotations at scale, the process was constantly monitored by the annotator project lead. In addition, to support annotators during the process, the UI allowed them to write comments in case a clarification was needed. Each comment was reviewed by the project lead, which reported to us any problem found during the process. We requested a re-annotation of the pairs in case of errors.

A.6 Annotation Efficiency

We investigate on the efficiency of this approach in terms of annotation cost. After sampling questions from existing datasets, we pair them with



(a)



(b)

Figure 4: Screenshot of the annotation tooling for (a) Question/Context validation and (b) answer collection.

context that are retrieved using the procedure described in Section 4.1. This procedure can lead to invalid pairs, which are discarded with a manual annotation. From the human annotations, we observed that: (i) 13% of the questions are discarded because *context-independent*; (ii) and 18% of the ⟨*question, context*⟩ pairs are discarded because *context-connected* or *context-answered*. However, based on the quote provided by our annotation provider to manually formulate questions, we found that the overall annotation cost is 31% lower when using our approach. This is because the manual question sourcing phase is more expensive than validating ⟨*question, context*⟩ pairs.

B Model Implementation Details

Question Rewriting. Question rewriting is a popular approach in Conversational QA to directly encode relevant information into the question without explicitly use the contextual models. To obtain the rewrites, we implemented the generative model proposed in (Anantha et al., 2021) that rewrites using the concatenation of the question and the *context*. We use a T5-base model available in HuggingFace (Wolf et al., 2020) and we trained it on the QReCC dataset with 5 epochs, a batch-size of 4, 500 warm-up steps, and a learning rate of $5e-5$.

AS2 and CS2 Model Training. Since ASNQ consists of open domain questions associated to a single Wikipedia page, only answer-level contexts can be extracted (document title, document content, or local context). In the case of *No-Context*, *Local*, and READOUT$_{\text{A}}$ models, the same context can be tuned on ASNQ first and FOCUSQA subsequently. The other contexts cannot be directly trained on ASNQ as the dataset does not contain that type of contextual information. In order to alleviate this issue we adopted models with conceptually similar contexts when adapting from ASNQ to the target domain. We adapted READOUT$_{\text{Q}}$ from READOUT$_{\text{A}}$, Cross-Context$_{\text{Titles}}$ from *Local*, and Cross-Context$_{\text{QA}}$ from a standard non contextualized model. We framed these contexts into the CS2 framework[5] proposed in Lauriola and Moschitti (2021). Similarly to existing CS2 solutions, these models are implemented through a multi-sentence Transformer that uses multiple token-type embeddings for each encoded text.

During the two fine-tuning stages, models were trained with (i) binary cross-entropy loss, (ii) Adam optimizer, (iii) batch-size of 768, (iv) triangular learning-rate scheduler whose peak comes after 0.15 epoch, (v) 15 maximum epochs. The development set was used to early stop the training when observing a decrease of P@1 after two consecutive epochs and to select the learning rate, with values $\{1, 5\} \cdot 10^{\{-6,-5\}}$. We set a max length of 256 for *No-Context* and *Local*, 320 for READOUT, and 512 for Cross-Context.

Starting from checkpoints trained on ASNQ, we repeated the fine-tuning (and model selection) step on FOCUSQA 3 times with 3 different random seeds. Eventually, we collected average and standard deviation of P@1 and other metrics. The training on ASNQ was done once per context type due to the dimension of the dataset and the associated training cost.

[5] https://github.com/alexa/wqa-contextual-qa