On the portability of extractive Question-Answering systems on scientific papers to real-life application scenarios

Chyrine Tahri [♣,♢] **Xavier Tannier** [♣] **Patrick Haouat** [♢]
[♣] Sorbonne Université, Inserm, Université Sorbonne Paris-Nord, LIMICS, Paris, France
[♢] ERDYN, Paris, France
{chyrine.tahri, xavier.tannier}@sorbonne-universite.fr
patrick.haouat@erdyn.fr

Abstract

There are still hurdles standing in the way of faster and more efficient knowledge consumption in industrial environments seeking to foster innovation. In this work, we address the portability of extractive Question Answering systems from academic spheres to industries basing their decisions on thorough scientific papers analysis. Keeping in mind that such industrial contexts often lack high-quality data to develop their own QA systems, we illustrate the misalignment between application requirements and cost sensitivity of such industries and some widespread practices tackling the domain-adaptation problem in the academic world. Through a series of extractive QA experiments on QASPER, we adopt the pipeline-based retriever-ranker-reader architecture for answering a question on a scientific paper and show the impact of modeling choices in different stages on the quality of answer prediction. We thus provide a characterization of practical aspects of real-life application scenarios and notice that appropriate trade-offs can be efficient and add value in those industrial environments.

1 Introduction

It is widely recognized today that the most advanced countries have moved to the so-called knowledge-based economy. In the industrial field, including service providers, this new paradigm has particular consequences for most players in R&D and innovation activities where decisions are based on the analysis of huge corpora of documents (scientific papers, patents, reports, etc). The thorough exploitation of this pre-existing knowledge by highly-skilled workers is costly and time-consuming, but such costs can be significantly reduced by NLP technologies that make exploitation and consumption of textual content faster and more efficient. For instance, Information-Seeking Question-Answering is of particular interest to industrial environments conducting scientific monitoring, but there still remain significant hurdles to efficiently adopt such systems in those environments, predominantly the complexity and accessibility of the data landscape.

As a matter of fact, extracting information from scientific publications is a cognitively complex process and requires domain expertise, but obtaining and ensuring such high-quality annotations could become unreasonably expensive and unreliable. The *scarcity of in-house annotation efforts*, *frequent domain shifts*, and *lack of deep understanding of data-model interaction and evaluation* make these technologies inaccessible especially for industrial environments lacking computational resources. One direction would be to entirely rely on models' transfer learning capabilities and make use of the knowledge they learn on academic benchmarks that meet the size requirement. However, zero and few-shot settings successes, *i.e,* when few to no annotations are available, seem to be largely dominated by large-scale autoregressive models (Chowdhery et al., 2022), which are accessible only to a handful of researchers and practitioners with enormous compute power.

In this paper, we take on extractive information-seeking QA on scientific papers from an industrial point of view. We identify the hurdles standing in the way of adopting such systems and show through a simulation of such context that some modeling and evaluation practices might not align with a suitable return on investment sought by such industries. Our contributions can be summarized as follows: First, we explore the portability challenges of QA models toward scientific content-consuming industrial environments and split them into three major long-standing issues. Second, we simulate through a series of experiments on QASPER (Dasigi et al., 2021) the context where information is sought in research papers and thus illustrate the identified portability issues. Third, we discuss based on the results the relevance of modeling and evaluation choices when compared to the goal of adequately

Proceedings of the 1st Workshop on Information Extraction from Scientific Publications, pages 67–77

solving the task in a cost-effective way.

2 The portability challenge in industrial environments

For small and medium-sized enterprises (SMEs) interested in Information-Seeking QA on scientific publications, the question of work to be done compared to the benefit of it is very important as it informs the way resources are allocated. When bringing advancements like QA systems into real-world applications suffering data scarcity issues, choosing a benchmark representative of contexts, questions, and answers one would expect in their application remains the most widely adopted practice for maximizing accuracy. Unfortunately, due to the fact that meeting an information need is a hard concept to quantify, adopting such technologies can fall short of quantitatively measuring the impact and the business value created. We discuss hereafter three major inter-connected long-standing issues that restrain from successful portability:

Issue 1: Modeling real-world problems is challenging. Question Answering aims at meeting an information need and providing a user with relevant answers to their questions. However, in domains with high levels of expertise, assisting professionals in such complex processes requires, depending on the nature of the query, cognitive abilities that AI systems have not yet matured to (Chollet, 2019). The AI community has factually been benchmarking intelligence by comparing the defined skill exhibited by AI and humans at specific tasks, and building special-purpose systems capable of handling narrow, well-described tasks, more and more above human-level performance. This created a plethora of QA benchmarks/tasks measuring very specific skills (Rogers et al., 2021) as opposed to the complex processes one would long for in intelligent systems. Further, annotating the required amount of quality data to build such systems can be unaffordable for many industries and organizations. The question that arises here is whether to favor quantity in task format adequacy and thus potentially model performance, or limited content representativeness with complexity that guarantees quality and better alignment with real-world applications.

Issue 2: There is a real need for transparency and confidence not only in predictions but also in the whole predictive process in a way that allows users to assess how well-informed their decisions would be. However, there still remains insufficient understanding of the capabilities and limitations of models and the way they interact with data during the different stages of their training (Ramnath et al., 2020; Zhou and Srikumar, 2021). Up until recently, there has been little guidance on the suitability of which models for which cases in Question Answering (Luo et al., 2022). Tremendous work continues to be done on modeling and exploring new model architectures and training schemes, however interpretation and explanation of models' behaviors that inform modeling choices in adopting such technologies, have not developed at the same pace. This makes it challenging for adopters to select their models for real-world settings, whether the intended use is at early stages or later in production. The obvious issue here is to identify what makes a certain model a trustworthy fit for the project motivation rather than another.

Issue 3: A good performance metric is not synonymous with how well application requirements are met. While current evaluation schemes contribute to overly specializing solutions for performance benchmarks, adopters and end-users are not only more sensitive to the plus-value models provide, but also the costs of developing and deploying such systems. Extractive QA systems are mainly evaluated using the F-measure, but a token-overlap metric is not informative on how well the system is assisting the user and providing relevant answers. For this reason, misaligning what is measured and what is intended and desired might lead in certain cases to misallocating resources, and although progress has been made towards user-centered evaluation (Chen et al., 2022), real-world applications still have more complexity and demands whereas models' evaluation is lagging behind.

These issues impact different phases of the development cycle of QA systems in real-world expert applications. For instance, issue 1 impacts problem definition and adequate data collection, which are the backbone of the whole cycle. Issue 2 introduces hurdles to experiment design and model training, while issue 3 directly impacts evaluation and complicates the path to successful model deployment. In the rest of the paper, we simulate a scenario of seeking information in research papers and consider our end-user to be an expert in decision support based on scientific publications analysis. We particularly focus on the extractive QA setting where the goal is to provide the user

with answers to a particular question on a given paper. This translates to the following formulation of the issues mentioned above: 1. What kind of task and data should we use, given the complexity and the level of expertise present both in the questions and the context? 2. What models would solve this task and would interact well with such data, peculiarly since we need as much transparency as possible in the process of identifying the answers? 3. How does the performance of the chosen model on the chosen task and data reflect the return on investment for deploying such systems?

3 Related work

3.1 Information-Seeking Question Answering

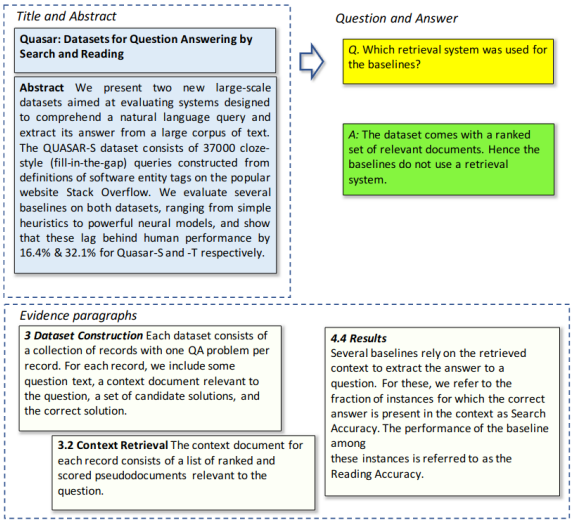

Figure 1: Example instance taken from QASPER as presented in Dasigi et al. (2021)

Rogers et al. (2021) make the distinction between information-seeking and probing questions based on the communicative intent of the user. We are more interested in information-seeking questions that aim to bring forth answers that are unknown at the time of formulating the query. There exists conversational information-seeking datasets such as QuAC (Choi et al., 2018), and grounded-in-documents datasets such as Natural Questions (Kwiatkowski et al., 2019) and QASPER (Dasigi et al., 2021).

3.2 Domain Adaptation for Question Answering

QA systems are often considered to be reliable when they have been trained on enough in-domain data, which is typically around 100k question and answer training examples. However, it is well known that such data is not abundant in specialized and restricted domains that require high-levels of expertise. Sparked industrial interest in QA use-cases has given rise to a line of work on Domain-Adaptation (Hazen et al., 2019; Miller et al., 2021; Yue et al., 2021) hoping to build robust systems for domains with limited data.

Overall, the general approach to domain adaptation of Question-Answering models is to synthesize question-answer pairs (Shinoda et al., 2021; Yue et al., 2022). Nevertheless, in the case of information-seeking QA on research papers, such approaches fall short of producing high-quality questions and are so far unable to efficiently deal with complex question and answer generation from long context dependencies (Luu et al., 2020). Therefore, domain adaptation of QA techniques cannot yet deal with generating synthetic, high-quality, and representative question-answer pairs of information sought in research papers.

3.3 Modular pipelined systems for Question Answering

Although modular pipelined QA systems are mainly developed and used in Open-Domain QA (Zhu et al., 2021), their components can be also beneficial for tackling in-context QA. Figure 2 shows the way we adopt retriever-ranker-reader architecture for answering a question on a scientific paper. We favor such building blocks of a solution rather than complex *do-it-all* models to increase our chances of understanding and trusting the system.

Retriever

A retriever aims at retrieving passages from a corpus that are relevant w.r.t. a given query. Its goal is to filter out irrelevant context and therefore it can be used in QA grounded in documents when these are very long sequences of text like research papers. The granularity of passages to be retrieved depends on the application and the type of answers sought.

State-of-the-art retrievers are mostly dense retrievers (Luan et al., 2021), *i.e,* they extract dense representations of a question and a context by feeding them into a language model and using the dot-product of these representations as a similarity score to rank and select most relevant passages.

Re-ranker

In information-seeking QA, especially on research papers, the end-user might not always employ the

terms in their query as they appear in context, whether for lexical reasons like specific terminology or simply because the terms themselves are sought by the query. To this end, in order to improve retrieval quality, a common strategy is to process the retrieved passages or answers using a re-ranking module. Rankers post-retrieval in particular are useful when retrievers have a high recall but fail to rank documents according to relevance, sometimes due to the semantic similarity between questions and passages being very low (Lin et al., 2020).

Reader

A reader infers the answer to the question from a set of ordered documents it receives in a pipelined QA system. Readers are generally regarded as either extractive or generative. Extractive readers mainly assume the correct answer is present in the context and usually focus on learning to predict the start and end position of the answer, while the generative ones generate the answers from their vocabulary. The choice of reader type depends on the nature of questions and context and therefore evaluation procedures differ (Zhang et al., 2020).

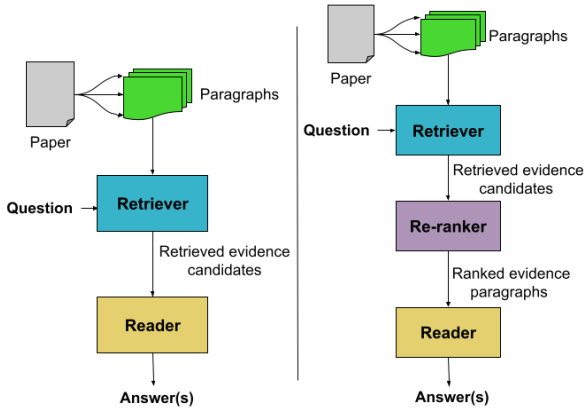

Figure 2: Modular pipeline for Information-Seeking Question Answering grounded in research papers. Left is a retriever-reader pipeline (referred to as pipeline **R**); Right is retriever-ranker-reader (**R-2**).

4 Experimental Setup

4.1 Datasets

QASPER for simulation

In restricted domains with high level of expertise, users tend to ask questions that are naturally different from those in open and general domains. For instance, the distribution of Google Search queries is not representative of all questions an astrophysicist or an economist routinely ask in a work-day. Such big datasets, arising from real-world use cases, might contain microscopic fractions of those specialized distributions one seeks, but will not be representative if regarded as a whole general domain. Therefore, we chose to focus our simulation on a dataset that drifts away from those general and "natural" distributions. To this end, QASPER (Dasigi et al., 2021) is an information-seeking dataset of questions and answers anchored in research papers whose main topic is NLP: it compromises 5,049 questions over 1,585 papers. The dataset is challenging in nature because of the long context requiring reading entire papers and the multiple types of questions (extractive, abstractive, yes/no, and unanswerable). Its task is formally defined as determining the answerability of the question and outputting an answer that can have different formats (span(s), free-form, yes/no).

We consider QASPER to be a good dataset for simulating an industrial environment seeking information in scientific text as the nature of the context, as well as the annotation strategy, are suitable and equivalent to our use-case. At the time of writing, it is currently the only existing benchmark focusing on entire research papers and not just abstracts.

The official baseline for QASPER is Longformer Encoder-Decoder (LED) (Beltagy et al., 2020). LED was trained in a multi-task setup for evidence identification and answer generation and chosen because of its ability to handle the variety of answer types as well as encoding papers' full text.

SQuAD

The Stanford Question Answering Dataset (Rajpurkar et al., 2016, 2018) has been widely used in QA tasks since its creation. It compromises over $100k$ crowd-sourced question-answer pairs derived from Wikipedia. Questions in SQuAD are diverse but answers are very short spans and require less expertise than QASPER to produce.

Natural Questions

Natural Questions (Kwiatkowski et al., 2019) introduced user queries issued to the Google search engine paired with high-quality annotations in the form of (*question, Wikipedia page, long answer, short answer*) quadruples. Additionally, Natural Questions is compromised of 323k examples, making it 64 times the size of QASPER.

4.2 Evidence Retrieval

For identifying relevant evidence paragraphs, we use Dense Passage Retrieval (DPR) (Karpukhin et al., 2020), a highly efficient dual-encoder using two BERT (Devlin et al., 2019) based models to encode documents and queries separately. Both the question encoder and context encoder have been trained on Natural Questions (Kwiatkowski et al., 2019). We use Haystack[1] as a framework and retrieval is performed using ElasticSearch.

Instead of encoding the entire long-context papers that cannot be handled with BERT-like encoders, and building on the definition of the task itself, *i.e,* identifying evidence paragraphs, we chose to deal with paragraphs as units of passages (Figure 2). Furthermore, 55% of the answers to questions with text-only evidence in QASPER have multiple evidence paragraphs. For this reason, and because retriever results could serve as explanations for the end-user and thus increase their confidence in predictions, we experiment with returning k candidate paragraphs with $k \in \{1, 3, 5, 10\}$. We chose these values to be "human-readable": an end-user is not visually bothered by having such $k \geq 1$ returned paragraphs highlighting answer elements.

Finally, because the semantic similarity between questions and passages can be very low (Figure 1), we experiment with re-ranking paragraphs using cross-encoders (Hofstätter et al., 2020) based on two models: MiniLM (Wang et al., 2020) and ELECTRA (Clark et al., 2020), trained on the MS Marco Passage Ranking[2] (Microsoft Machine Reading Comprehension) task. We choose to pass the minimum between the top-50 ranked paragraphs and the total number of paragraphs in the article[3] to the re-ranker because of its computational cost.

4.3 Answer Prediction

QASPER is composed of questions with multiple evidence and answer types. We focus on text-only evidence excluding tables and figures. We further limit experiments to extractive questions as we mentioned before (roughly 51.8% of the dataset) because we prioritized our focus on accessible and extensively-studied models as well as the extractive evaluation scheme. Finally, because we use a pipelined system with paragraphs as units of passages, we are able to fit candidate evidence in all readers[4]. We conduct two sets of experiments:

- Zero-shot settings on a few selected models that are known for robustness, generalization ability, and efficiency among others. This scenario is the closest to a real-world setting where no annotated data is available and the application is quite different from existing benchmarks. Such experiments lay the ground for what can be expected in a least-available resources scenario and it is interesting to see if there is value in those settings.

- Fine-tuned settings where all models are fine-tuned on the extractive set of questions in QASPER. We are particularly interested in seeing how models adapt their answers to better suit the answers' nature in QASPER. Since there would intuitively be improvements over the zero-shot setting when fine-tuning, this kind of scenario gives hints about the relevance of investing in expert annotations when considering the nature of such improvements.

The readers we chose to experiment with are the following: **RoBERTa** (Liu et al., 2019) offering a great trade-off between performance and inference speed, **SciBERT** (Beltagy et al., 2019) trained on scientific text, **deBERTaV3** (He et al., 2021) particularly performing on NLU tasks, **UnifiedQA** (Khashabi et al., 2020) for its strong generalization abilities and **Longformer** (Beltagy et al., 2020) which, although we do not need long-range models as the pipeline deals with paragraphs as units, has the ability to produce longer answer spans if needed.

We choose to have RoBERTa, SciBERT, deBERTa and Longformer trained on SQuAD v2.0 (Rajpurkar et al., 2018) because it is a simple and accessible starting point, *i.e,* a widely used dataset and trained models are open-sourced. UnifiedQA has been trained on other datasets with other formats in addition to SQuAD.

5 Results

We present in this section the results of the different stages of the pipeline when adding components or using different training strategies.

[1] https://github.com/deepset-ai/haystack

[2] https://github.com/microsoft/MSMARCO-Passage-Ranking

[3] Articles in QASPER have a number of paragraphs ranging from ≈ 20 to a maximum of ≈ 230

[4] For readers with 512 tokens limit, one passage exceeded the maximum length so we truncated the input.

		Evidence Span (F_1)				Top-k retrieval accuracy (%)			
LED		32.28				-			
Retriever ↓	Ranker ↓	k = 1	k = 3	k = 5	k = 10	k = 1	k = 3	k = 5	k = 10
DPR	w/o	37.68	54.73	66.68	79.38	23.23	40.69	55.57	71.86
	ELECTRA	52.63	72.07	80.28	89.17	39.08	62.63	73.45	85.60
	MiniLM	**54.65**	**74.05**	**81.76**	**90.91**	**41.54**	**65.31**	**75.48**	**87.97**

Table 1: Evidence F_1 and top-k retrieval accuracy on extractive questions in QASPER test.

5.1 Evidence retrieval

We show in Table 1 the results of the evidence retrieval stage with and without the use of a re-ranker for $k \in \{1, 3, 5, 10\}$ where k is the number of retrieved paragraphs. For $k > 1$, evidence-span (F_1) refers to the maximum overlap found between the gold evidence and the k retrieved paragraphs, whereas top-k retrieval accuracy (%) considers the case where an exact match is found within the top-k retrieved elements. We chose to report this metric because it is more informative to the end-user.

The retriever adequately improves with greater values of k, which is expected since the more it retrieves the more chances of finding a relevant paragraph. However, the use of the re-ranker considerably enhances the evidence retrieval step, with an average gain of $13.92F_1$ points with ELECTRA, and $15.73F_1$ points with MiniLM for the different values of k. In terms of retrieval accuracy, re-ranking adds on average 17.35% accuracy with ELECTRA and 19.74% with MiniLM. If we want to avoid overloading the end-user with irrelevant/incomplete evidence, using a ranker with a smaller k can be a very good option.

5.2 Answer identification

We select the best performing retrieval pipeline, i.e, DPR and MiniLM, and test different readers for end-to-end answer selection. We report the results in Table 2: for pipelines where $k > 1$, the reader produces an answer a_i for each retrieved (ranked) paragraph p_i. The results show the maximum overlap between $\{a_i\}_{i \leq k}$ and gold answers[5].

In both zero-shot and fine-tuned settings, all models surpass the LED baseline when returning $k \geq 3$ with ranking (note that LED does not return multiple candidates). When seeing QASPER for the first time, deBERTa outperforms the rest of the models, widening the gap with greater values of k. It is interesting to see that RoBERTa, UnifiedQA, Longformer and SciBERT have very close scores to each other.

Further, finetuning on QASPER does not preserve the performance ranking of models: UnifiedQA outperforms $\forall k \in \{1, 3, 5, 10\}$ all other models, both with and without ranking. This is to be expected with such generalization abilities. Unsurprisingly, models do not all benefit the same from re-ranking and fine-tuning as discussed in Issue 2. We present hereafter the differences in end-to-end performance gain for each model.

Effect of re-ranking

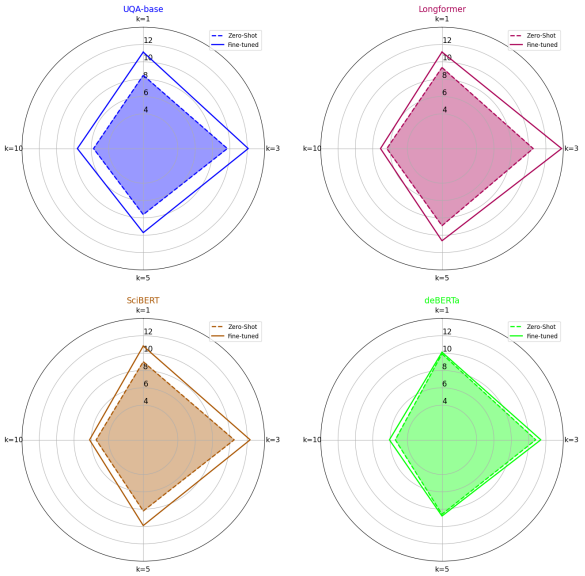

Figure 3: Gain in Answer-Span (F_1) when reranking

Figure 3 shows how much performance UnifiedQA, Longformer, SciBERT and deBERTa gain from re-ranking. The cross-encoder pre-reading helps improve answer identification in all scenarios: $\forall k \in \{1, 3, 5, 10\}$, with and without fine-tuning. The most significant gains are observed for $k = 1$ (a model average of $9.2F_1$(*zero-shot*) and $10.73F_1$(*fine-tuned*)) and $k = 3$ ($10.45F_1$ and $12.33F_1$ respectively). This is a sign of the ranker

[5]In QASPER, many questions have multiple annotators and therefore many answers. In v0.3, the answers have the same nature, *i.e,* all extractive in our case.

	Answer-span (F_1) end-to-end							
LED	32.0 (29.97*)							
DocHopper	36.4$^\diamond$							
	k = 1		k = 3		k = 5		k = 10	
	R	R-2	R	R-2	R	R-2	R	R-2
RoBERTa-base	13.01	22.08	25.05	35.75	32.28	40.52	40.10	45.82
UnifiedQA-base	13.58	22.03	25.31	35.02	32.47	40.09	40.61	46.35
Longformer	11.96	21.49	24.59	35.12	31.30	40.20	39.36	45.73
SciBERT	13.23	22.24	25.24	35.74	32.21	40.49	40.89	46.33
deBERTa	12.87	*22.79*	25.70	**36.53**	33.57	**42.15**	42.76	**48.15**
RoBERTa-base$_{ft}$	15.57	26.00	28.42	40.37	36.58	45.62	45.59	51.52
UnifiedQA-base$_{ft}$	16.41	*27.54*	30.30	**42.42**	38.14	**47.84**	47.47	**55.08**
Longformer$_{ft}$	15.66	26.80	28.32	42.13	36.58	47.22	45.60	52.70
SciBERT$_{ft}$	15.80	26.62	28.79	41.13	36.71	46.60	46.42	52.62
deBERTa$_{ft}$	16.34	26.45	30.01	41.42	38.14	46.87	47.12	53.19

Table 2: Answer-span predictions on extractive questions in QASPER test using DPR and MiniLM for retrieval. ft stands for further fine-tuning on QASPER. (* reported in Dasigi et al. (2021), $^\diamond$ reported in Sun et al. (2021))

propelling better context at the top. For all values of k, Longformer benefits most from re-ranking.

Effect of fine-tuning

Similarly, Figure 4 shows the gain in performance that the two pipelines benefit from when fine-tuning readers on QASPER. In all scenarios, fine-tuning enhances performance, with UnifiedQA having the largest gains (an average of $4F1$(*without-ranking*) and $7.35F1$(*with-ranking*)). The greater the value of k, the more models benefit from fine-tuning. This is due to the retrieval stage providing more relevant context.

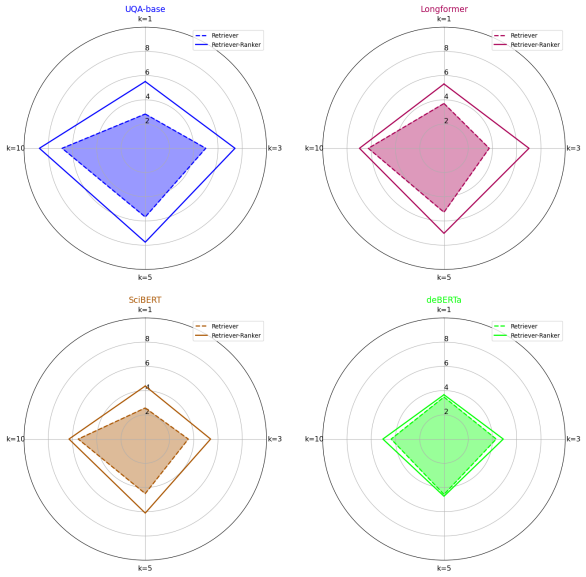

Figure 4: Gain in Answer-Span (F_1) when fine-tuning

6 Discussion

We discuss hereafter the sources of improvements and their alignment with the portability challenges.

6.1 Retrieval stage

Intuitively, we suspect that LED is under-optimized not only due to the size of QASPER but also because it treats evidence selection as a classification task (which is probably good for dealing with multiple evidence). DPR on the other hand has an appropriate training dataset size and the approach is done in a contrastive learning setting that might be better aligned with identifying how close a passage is to a query. If we consider the modeling issue (1) in this case, Natural Questions is one of the benchmarks that have the most similarities with QASPER: different levels of granularity (long and short answers), different types of answers, and no observation bias. With DPR being trained on NQ, this offers an adequate trade-off between task format adequacy and content representativeness.

The remaining issue would be the low semantic similarity faced in Information-Seeking QA grounded in research papers, which we try to circumvent with the use of a re-ranker. The latter very significantly enhances the evidence selection stage. As research papers are themed and specific, and throughout an article, information is redundant with varying degrees of detail, a bi-encoder might not be enough to relevantly score those differences. Additionally, top-k retrieval accuracy is

Paper	Question	Zero-Shot	Fine-Tuned	Gold Answer
1602.03483	Which unsupervised representation-learning objectives do they introduce?	Sequential Denoising Autoencoders	Sequential Denoising Autoencoders (SDAEs) and FastSent, a sentence-level log-linear bag-of-words model	Sequential Denoising Autoencoders (SDAEs) and FastSent
1606.07043	On which corpora do they evaluate on?	20 Newsgroup	20 Newsgroups and the i2b2 2008 Obesity Challenge BIBREF22 data set	20 Newsgroups, i2b2 2008 Obesity Challenge BIBREF22 data set
1602.04341	What was the margin their system outperformed previous ones?	15.6/16.5	The margin between our best-performing ABHCNN-TE and NR is 15.6/16.5 (accuracy/NDCG) on MCTest-150 and 7.3/4.6 on MCTest-500	15.6/16.5 (accuracy/NDCG) on MCTest-150 and 7.3/4.6 on MCTest-500
1707.07212	What are the components of the classifier?	context words, distance between entities	context words, distance between entities, presence of punctuation, dependency paths, and negated keyword	log-linear model, five feature templates: context words, distance between entities, presence of punctuation, dependency paths, and negated keyword

Table 3: Longformer's predictions where the fine-tuned model produces longer spans over the zero-shot prediction.

more informative than span-F_1: for instance with an appropriate retriever and ranker, the user can expect to have 3 questions over 4 where a correct evidence paragraph is placed within 5 suggestions.

6.2 Reading stage

Having models that are fine-tuned with large general-domain datasets before fine-tuning on QASPER is helpful. However, It has to be kept in mind that higher performance is not necessarily a sign of different and thus better answer identification, as the $F1$ metric does not faithfully reflect the actual performance (especially if the difference is about very few points): greater (lesser) non-zero values of $F1$ are not systematic indicators of better (worse) candidate answers (Bulian et al., 2022). The fact that many models have extremely small differences of performance in zero-shot emphasizes the need to look for other preferences than performance when selecting readers before considering investing in their improvement; for instance an ability to return longer answers. To this end, we examined Longformer's predictions in the case $k = 1$, *i.e,* either it receives correct evidence or not, to see how faithful the performance gain is to the improvement of predictions. When investigating the questions where fine-tuning improved the zero-shot prediction, we surprisingly noticed that the gained performance in the pipeline R is due in 36.36% of cases to longer answers containing the string of the zero-shot prediction. Similarly for pipeline R-2, 43.78% of the improved answers are merely longer spans. This might be a sign of completeness, but how necessary is it really compared to the cost of attaining such gains if the answer is visually located in its context? We provide examples of such predictions in Table 3.

6.3 Implications for the portability issues

In real-world settings, a user seeking information in scientific publications might face very frequent topic change. It is well known, both in academia and industry, that QA annotations on scientific papers is extremely scarce: QASPER is the current only benchmark on entire papers. Further, its subset of extractive questions compromises over 1000 expert-annotated questions. As this is very expensive to obtain, users will be tempted to focus on

zero-shot settings performance. We discuss hereafter the implications for the portability issues from what we observed on QASPER:

Issue 1: Current benchmarks do not faithfully translate the complexity of tasks humans carry in their quests for innovation and knowledge consumption and there is a tendency to criticize how far real-world data can be from such datasets. Because obtaining high-quality and representative annotations in such environments is way too costly, there can be a plus-value in trading-off content representativeness with task format adequacy. For instance, Natural Questions accounts for a great "similar" task for the retrieval stage.

Issue 2: In some cases, accessible models trained on adequate benchmarks can provide satisfying zero-shot results without incurring the need to invest in having a greater reported F1. To this end, building simple and fast to deploy blocks of a solution does not imply jeopardizing performance since design complexity is not necessarily the ground-laying part of accuracy: LED is outperformed by simpler pipelines offering more transparency of the whole predictive process.

Issue 3: Users should align their application needs with models' characteristics rather than solely focusing on performance metrics and the processes of improving it. Not only enhancing model performance by fine-tuning on domain-specific data might not align well with the cost sensitivity of adopters, but also experts seeking to more efficiently consume scientific content are not to be withdrawn from the information-seeking process the greater the reported performance metric is. For instance, a user visually locating the answer span in a paper accounts for 43% of Longformer's performance improvement with fine-tuning (and the related costs).

6.4 Limitations of this work

We did not experiment on few-shot settings, eventhough such scenarios are anchored in real-world settings. The reason for this is that such scenarios heavily rely on data augmentation techniques; but these approaches fall short of producing the quality we seek in such annotations as we explained in Section 3.2. Therefore we are left with large autoregressive models with stunning few-shot abilities, but those are not yet accessible options either. Another limit is that we restrained our experiments to the extractive questions only. We made this choice because evaluation schemes would be more complex and it would be harder to interpret performance variations (Gehrmann et al., 2022). It is also not mandatory from the industrial point of view at this time to go beyond extractive models, as these already have a plus-value for the workers.

7 Conclusion

Information-seeking QA on scientific content is gaining popularity in a world of knowledge-based economies. In this paper, we identified the hurdles that stand in the way of efficient portability of such systems into industrial environments suffering data scarcity. We revealed through a series of experiments on extractive QA anchored in research papers, that bridging the gap between academic benchmarks along with their models' performance, and concrete user needs that are most often hindered by resource allocation constraints in business can be done with appropriate trade-offs and that caution needs be taken when investing in widespread but costly practices.

Acknowledgements

This work has been funded by the ANRT CIFRE convention N°2019/1314 and ERDYN. We would like to thank the reviewers for their valuable and insightful comments.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *EMNLP*. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. *CoRR*, abs/2202.07654.

Yuyan Chen, Yanghua Xiao, and Bang Liu. 2022. Grow-and-clip: Informative-yet-concise evidence distillation for answer explanation.

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, Brussels, Belgium. Association for Computational Linguistics.

François Chollet. 2019. On the measure of intelligence.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4599–4610. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text.

Timothy J. Hazen, Shehzaad Dhuliawala, and Daniel Boies. 2019. Towards domain adaptation from limited data for question answering using deep neural networks.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained transformers for text ranking: Bert and beyond.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Man Luo, Kazuma Hashimoto, Semih Yavuz, Zhiwei Liu, Chitta Baral, and Yingbo Zhou. 2022. Choose your QA model wisely: A systematic study of generative and extractive readers for question answering. In *Proceedings of the 1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge*, pages 7–22, Dublin, Ireland and Online. Association for Computational Linguistics.

Anh Tuan Luu, Darsh J. Shah, and Regina Barzilay. 2020. Capturing greater context for question generation. In *AAAI*.

Timothy Miller, Egoitz Laparra, and Steven Bethard. 2021. Domain adaptation in practice: Lessons from a real-world information extraction pipeline. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 105–110, Kyiv, Ukraine. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for 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.

Sahana Ramnath, Preksha Nema, Deep Sahni, and Mitesh M. Khapra. 2020. Towards interpreting BERT for reading comprehension based QA. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3236–3242, Online. Association for Computational Linguistics.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2021. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension.

Kazutoshi Shinoda, Saku Sugawara, and Akiko Aizawa. 2021. Improving the robustness of QA models to challenge sets with variational question-answer pair generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 197–214, Online. Association for Computational Linguistics.

Haitian Sun, William W. Cohen, and Ruslan Salakhutdinov. 2021. Iterative hierarchical attention for answering complex questions over long documents.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Xiang Yue, Ziyu Yao, and Huan Sun. 2022. Synthetic question value estimation for domain adaptation of question answering.

Zhenrui Yue, Bernhard Kratzwald, and Stefan Feuerriegel. 2021. Contrastive domain adaptation for question answering using limited text corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zhuosheng Zhang, Hai Zhao, and Rui Wang. 2020. Machine reading comprehension: The role of contextualized language models and beyond. *ArXiv*, abs/2005.06249.

Yichu Zhou and Vivek Srikumar. 2021. A closer look at how fine-tuning changes bert.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering.