FQuAD2.0: French Question Answering and Learning When You Don't Know

Quentin Heinrich, Gautier Viaud, Wacim Belblidia
Illuin Technology
Paris, France
{quentin.heinrich, gautier.viaud, wacim.belblidia}@illuin.tech

Abstract

Question Answering, including Reading Comprehension, is one of the NLP research areas that has seen significant scientific breakthroughs over the past few years, thanks to the concomitant advances in Language Modeling. Most of these breakthroughs, however, are centered on the English language. As a first strong initiative to bridge the gap with the French language, FQuAD1.1 was introduced in 2020: it is a French Native Reading Comprehension dataset composed of 60,000+ questions and answers samples extracted from Wikipedia articles. Nonetheless, Question Answering models trained on this dataset have a major drawback: they are not able to predict when a given question has no answer in the paragraph of interest, therefore making unreliable predictions in various industrial use-cases. We introduce FQuAD2.0, which extends FQuAD with 17,000+ unanswerable questions, annotated adversarially, in order to be similar to answerable ones. This new dataset, comprising a total of almost 80,000 questions, makes it possible to train French Question Answering models with the ability of distinguishing unanswerable questions from answerable ones. We benchmark several models with this dataset: our best model, a fine-tuned CamemBERT$_{LARGE}$, achieves a F1 score of 82.3% on this classification task, and a F1 score of 83% on the Reading Comprehension task.

Keywords: question answering, adversarial, french, multilingual

1. Introduction

Question Answering (QA) is a central task in Natural Language Understanding (NLU), with numerous industrial applications such as searching information in large corpus, extracting information from conversations or form filling. Amongst this domain, Reading Comprehension has gained a lot of traction in the past years thanks to two main factors. First, the release of numerous datasets such as SQuAD1.1 (Rajpurkar et al., 2016), SQuAD2.0 (Rajpurkar et al., 2018), BoolQ (Clark et al., 2019), CoQA (Reddy et al., 2018), Natural Questions (Kwiatkowski et al., 2019), only to cite a few. Second, the progress in Language Modeling with the introduction of transformers model (Vaswani et al., 2017), leveraging self-supervised training on very large text corpus, followed by fine-tuning on a downstream task (Devlin et al., 2018). This process has now become a de-facto standard for most of Natural Language Processing (NLP) tasks and also contributed to the progress of state-of-the-art for most of these tasks, including Question Answering.

Whilst most of these recent transformers models are English models, French language models have also been released, in particular CamemBERT (Martin et al., 2019) and FlauBERT (Le et al., 2019), as well as multilingual models such as mBERT (Pires et al., 2019) or XLM-RoBERTa (Conneau et al., 2019). These models fostered the state-of-the-art in NLP for French language, allowing to benefit from this large-scale transfer learning mechanism. However, native French resources for Question Answering remain scarcer than English resources. Nonetheless, in 2020, FQuAD1.1

(d'Hoffschmidt et al., 2020) was introduced: with 60,000+ question-answer pairs, it enabled the development of QA models surpassing the human performance in the Reading Comprehension task.

Although, the introduction of this resource was a major leap forward for French QA, the obtained models suffered from an important weakness, as they were trained to consistently find an answer to the specified question by reading the associated context. In real-life applications however, it is often the case that asked questions do not have an answer in the associated context. For example, let us imagine that we are building a system designed to automatically fill a form with relevant information from property advertisements. We could be interested in the type of property, its surface area and its number of rooms. By asking questions such as *"How many rooms does the property have?"* to a QA model, we would be able to extract this information. But it would also be important that our model is able to predict when such a question does not find an answer in the provided text advertisement, as this situation often arises.

As FQuAD1.1 contains solely answerable questions, the models trained on this dataset did not learn to determine when a question is unanswerable with the associated context. To overcome this difficulty, we extended FQuAD with unanswerable questions, annotated adversarially, in order to be close to answerable ones.

Our contribution sums as follows:

- We introduce the FQuAD2.0 dataset, which extends FQuAD1.1 with 17,000+ unanswerable questions, hand-crafted to be difficult to dis-

tinguish from answerable questions, making it the first French adversarial Question Answering dataset with a grand total of almost 80,000 questions.

- We evaluate how models benefit from being trained on adversarial questions to learn when questions are unanswerable. To do so, we fine-tune CamemBERT models of varying sizes (large, base) on the training set of FQuAD2.0 and evaluate it on the development set of FQuAD2.0. We take interest in both the ability of a model to distinguish unanswerable questions from answerable ones, as well as an eventual performance drop in the precision of answers provided for answerable questions, due to the addition of unanswerable questions during training. We also study the impact of the number of adversarial questions used and obtain learning curves for each model.

- By using both FQuAD2.0 and SQuAD2.0 datasets, we study how multilingual models fine-tuned solely on question-answer pairs of a single language (English), performs in another language (French). We also take interest in performances of such models trained on both French and English datasets.

2. Related work

In the past few years, several initiatives emerged to promote Reading Comprehension in French. With both the release of large scale Reading Comprehension datasets in English such as SQuAD (Rajpurkar et al., 2016; Rajpurkar et al., 2018), and the drastic improvement of Neural Machine Translation with the emergence of the attention mechanism within models architectures (Bahdanau et al., 2016; Vaswani et al., 2017), it was at first the most natural path to try to machine translate such English datasets to French. This is what works such as (Asai et al., 2018) or (Kabbadj, 2018) experimented. However, translating a Reading Comprehension dataset presents inherent difficulties. Indeed, in some cases, the context translation can reformulate the answer such that it is not possible to match it to the answer translation, making the sample unusable. Specific translation methods to mitigate this difficulty were for example proposed in (Asai et al., 2018) or (Carrino et al., 2019a). Such translated datasets then enable the training of Question Answering models with French data. However, (d'Hoffschmidt et al., 2020) demonstrates that the use of French native datasets such as FQuAD1.1 brings far better models. In addition to FQuAD, another French native Question Answering dataset has been released: PiAF (Rachel et al., 2020). This smaller dataset, complementary to FQuAD1.1, contains 3835 question-answer pairs in its 1.0 version and up to 9225 question-answer pairs in its 1.2 version. We described in Section 1 the purpose of adversarial Question Answering datasets for some use cases.

This concept was first introduced in (Rajpurkar et al., 2018) with the presentation of SQuAD2.0, an English dataset extending SQuAD1.1 (Rajpurkar et al., 2016) with over 50,000 unanswerable questions. The purpose of SQuAD2.0 is similar to FQuAD2.0's and the hereby presented work takes some of its roots in its English counterpart.

To the best of our knowledge, there is no other dataset for adversarial Question Answering other than SQuAD2.0 and FQuAD2.0, even though there exist several translated versions of SQuAD2.0 in Spanish (Carrino et al., 2019b), Swedish (Okazawa, 2021), Polish[1], Dutch[2], among others. This makes FQuAD2.0 the first non-English native dataset for adversarial QA.

Nonetheless, numerous large-scale Question Answering datasets exist apart from FQuAD2.0 and SQuAD2.0, focusing on different Question Answering paradigms. Some take interest in QA within conversations: CoQA (Reddy et al., 2018) highlights the difficulties of answering interconnected questions that appear in a conversation, QuAC (Choi et al., 2018) provides questions asked during an Information Seeking Dialog, ShARC (Saeidi et al., 2018) studies real-world scenarios where follow-up questions are asked to obtain further informations before answering the initial question. Others take interest in more complex forms of reasoning: HotpotQA (Yang et al., 2018) contains questions that require reasoning among multiple documents, DROP (Dua et al., 2019) introduces questions requiring discrete reasoning types such as addition, comparison or counting. MASH-QA (Zhu et al., 2020) takes interest in questions that have multiple and non-consecutive answers within a document. BoolQ (Clark et al., 2019) highlights the surprising difficulty of answering yes/no questions, while RACE (Lai et al., 2017) studies multiple choice Question Answering. The last paradigm we would like to mention is Open Domain Question Answering where a system is asked to find the answer of a given question among a large corpus. This task is tackled with datasets such as Natural Questions (Kwiatkowski et al., 2019) or TriviaQA (Joshi et al., 2017).

Another line of research we find very interesting is a model ability to learn to jointly tackle several Question Answering tasks. (Micheli et al., 2021) trained RoBERTa models (Liu et al., 2019) on both BoolQ (Clark et al., 2019), a boolean Question Answering dataset, and SQuAD2.0. To enable the model to be multi-task, a slightly modified RoBERTa architecture is presented. Another way to obtain multi-task Question Answering models is to use an autoregressive model such as GPT-3 (Brown et al., 2020) or T5 (Raffel et al., 2020). When provided with a context and a question, these text-to-text models directly output an answer in the form of a sequence of tokens using their decoder. This naturally enables such approaches to gener-

---

[1] https://bit.ly/2ZqLLgb
[2] https://bit.ly/3CHFyee

alize beyond Reading Comprehension. For example a boolean question can simply be treated by a text-to-text model by outputing "yes" or "no". (Khashabi et al., 2020) leveraged this concept even further by training a T5 model to tackle four different Question Answering formats: Extractive, Abstractive, Multiple Choice, and Boolean. Their model performs then well accross 20 different QA datasets. We believe that such models for the French language will soon appear, building on French datasets such as FQuAD2.0.

3. Dataset collection & analysis

3.1. Annotation process

FQuAD2.0 is an extension of FQuAD1.1 (d'Hoffschmidt et al., 2020). This extension consists in the addition of unanswerable questions. These questions are hand-crafted in an adversarial manner in order to be difficult to distinguish from answerable ones. To achieve this goal we gave precise guidelines to the annotators:

- An adversarial question must be relevant to the context paragraph by addressing a topic also addressed in the context paragraph.

- An adversarial question should be designed in the following way: ask an answerable question on the paragraph, and apply to it a transformation such as an entity swap, a negation or something else that renders the question unanswerable.

The articles and paragraphs used in the train, development and test sets of FQuAD1.1 and FQuAD2.0 are exactly the same. An annotator is presented with a paragraph and the already existing answerable questions collected for this paragraph for FQuAD1.1. He is then asked to forge at least 4 adversarial questions, while spending up to 7 minutes by paragraph. A total of 17,765 adversarial questions were collected in 3,100 paragraphs. As FQuAD contains in total 14,908 paragraphs, unanswerable questions were not annotated for every paragraph, nor every article. In order to have reliable evaluations on the development and test sets for this new task, we chose to annotate, in proportion, an important amount of adversarial questions in these two sets. They contain in total around 42% adversarial questions, while the train set contains 16% adversarial questions. More statistics can be found in table 1.

We used the Étiquette annotation platform[3] developed by Illuin Technology. It has a dedicated interface to annotate the Question Answering task. Unanswerable questions can be annotated by indicating that the answer is an empty string. A screenshot of the platform is displayed in Figure 1.

A total of 18 French students contributed to the annotation of the dataset. They were hired in collaboration with the Junior Enterprise of CentraleSupélec[4]. To

limit the bias introduced by an annotator's own style of forging adversarial questions, each annotator only contributed to a given subset: train, development or test.

3.2. Statistics

To maximize the number of available questions for fine-tuning experiments, while keeping a sufficiently important set for evaluation, we decide to merge the train and test sets of FQuAD2.0 into a bigger training set, and keep the development set intact for evaluating the obtained models. The new training set contains a total of 13,591 unanswerable questions. Main statistics for FQuAD1.1 and FQuAD2.0 are presented in Table 1.

3.3. Challenges raised by adversarial questions

To understand what the different types of adversarial questions collected are, we propose a segmentation of the challenges raised by adversarial questions in FQuAD2.0. To do so, we randomly sampled 102 questions from the new annotated questions in FQuAD2.0 development set and manually inspected them to identify the challenges they proposed. Then, we sorted these questions following the different identified categories, in order to estimate the proportion of each category within the total dataset. Table 2 presents this analysis where 5 main categories have been identified.

4. Evaluation metrics

To evaluate the predictions of a model on the FQuAD2.0 dataset, we use mainly the Exact Match (EM) and F1 score metrics, which are defined exactly as in (Rajpurkar et al., 2018) with the required adaptations regarding stop words for the French language as explained in (d'Hoffschmidt et al., 2020). In a nutshell, EM measures the percentage of predictions matching exactly one of the ground truth answers, while F1 computes the average overlap between the predicted tokens and the ground truth answer.

To extend these metrics to unanswerable metrics, unanswerable questions are simply considered as answerable questions with a ground truth answer being an empty string.

One may be interested in evaluating on the one hand the ability of a model to extract the correct answers of answerable questions, and on the other hand its ability to determine if a question is unanswerable given the context. To do so, we introduce two other metrics:

- $F1_{has\ ans}$: the average F1 score, question-wise, as defined above, but limited to answerable questions,

- $NoAns_{F1}$: the F1 score of the classification problem consisting in determining if a question is unanswerable. It is then the harmonic mean of the precision ($NoAns_P$) and recall ($NoAns_R$) for this classification problem, the no-answer class being considered as the positive class.

Create question/answer pairs for the text below
Avoid using the same words/sentences as in the text when asking a question. You are incentived to ask **difficult** questions.

La succession d'Assarhaddon, en 669 av. J.-C., avait en fait donné lieu à une organisation politique spéciale : Assurbanipal régnait depuis l'Assyrie, alors que son frère Shamash-shum-ukin était placé sur le trône de Babylone, en position de vassal mais auréolé du retour de la statue de Marduk qui accompagne son intronisation. Ce dernier se révolte finalement en 652, mais est vaincu après une guerre âpre de quatre ans et le siège de sa ville qui dure plusieurs mois en 648. Il meurt lors du siège de Babylone, brûlé dans l'incendie de son palais, histoire qui donna naissance au mythe grec de Sardanapale. Après une première phase de répression Assurbanipal se révèle moins brutal que son grand-père et fait restaurer la ville, à la tête de laquelle il place un souverain fantoche, Kandalanu. Finalement, les rois assyriens ont profondément marqué l'histoire de Babylone et sans doute aussi son paysage urbain.

Questions

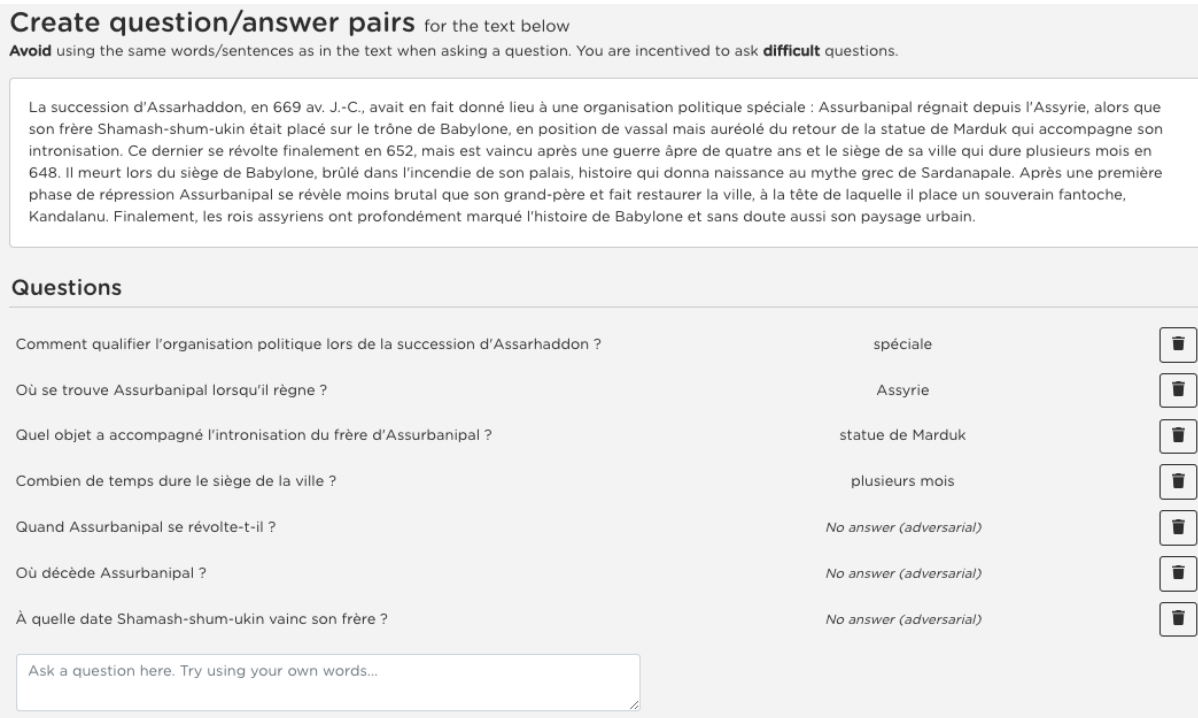|	
Comment qualifier l'organisation politique lors de la succession d'Assarhaddon ?	spéciale
Où se trouve Assurbanipal lorsqu'il règne ?	Assyrie
Quel objet a accompagné l'intronisation du frère d'Assurbanipal ?	statue de Marduk
Combien de temps dure le siège de la ville ?	plusieurs mois
Quand Assurbanipal se révolte-t-il ?	*No answer (adversarial)*
Où décède Assurbanipal ?	*No answer (adversarial)*
À quelle date Shamash-shum-ukin vainc son frère ?	*No answer (adversarial)*

Ask a question here. Try using your own words...

Figure 1: The interface used to collect the question-answer pairs for FQuAD. During the annotation process for FQuAD2.0, an annotator can see a paragraph and the associated answerable questions that were already collected for FQuAD1.1.

	FQuAD1.1			FQuAD2.0		
	Train	Development	Test	Train	Development	Test
Articles	271	30	25	271	30	25
Paragraphs	12,123	1,387	1,398	12,123	1,387	1,398
Answerable questions	50,741	5,668	5,594	50,741	5,668	5,594
Unanswerable questions	0	0	0	9,481	4,174	4,110
Total questions	50,741	5,668	5,594	60,222	9,842	9,704

Table 1: Dataset statistics for FQuAD1.1 and FQuAD2.0

We must emphasize that NoAns_{F1} is a metric computed as a whole on the entirety of the FQuAD2.0 development set as a classification problem, while $\text{F1}_{\text{has ans}}$ is computed question-wise and is an average of the individual scores for each question. We also want to point out that the global F1 score is in no way the weighted average of $\text{F1}_{\text{has ans}}$ and NoAns_{F1}.

In order to evaluate the ability of a Question Answering model to learn when a question is unanswerable, we carried out various fine-tuning experiments using the FQuAD2.0 dataset. These experiments are split into the following sections: French Monolingual Experiments and Multilingual Experiments.

5. French monolingual experiments

The goal of these experiments is two-fold. First, we want to obtain strong baselines on the FQuAD2.0 dataset. Second, we want to analyze how the performances of the fine-tuned models evolve with respect to the quantity of unanswerable questions available at training time.

5.1. Baselines

To fulfill our first goal, we choose to fine-tune Camem-BERT models, because they are the best performing models on several NER, NLI and Question Answering French benchmarks (Martin et al., 2019; d'Hoffschmidt et al., 2020). We could also have chosen FlauBERT models (Le et al., 2019), but (d'Hoffschmidt et al., 2020) tends to show that for the same size, CamemBERT models outperform FlauBERT models on the Question Answering task, hence our choice of CamemBERT models. We benchmark two different model sizes: CamemBERT$_{\text{LARGE}}$ (24 layers, 1024 hidden dimensions, 12 attention heads, 340M parameters) and CamemBERT$_{\text{BASE}}$ (12 layers, 768 hidden dimensions, 12 attention heads, 110M parameters).

The fine-tuning procedure used is identical to the one described in (Devlin et al., 2018), and an implementation can be found in HuggingFace's Transformers li-

Reasoning	Description	Example	Frequency
Antonym	Use of negation or antonym to make the question adversarial.	Question: Quels mamifères **ne** sont **pas** présents ? Context: [...] Le parc abrite aussi de nombreux grands mammifères comme **des ours noirs, des grizzlys**, [...]	21.6 %
Entity Swap	A name, a number, a date has been modified so that the question becomes adversarial.	Question: Quelle est la couleur traditionnelle de la ville de Paris ? Context: [...] La livrée des rames est personnalisée, associant **le vert jade** traditionnel de **la RATP** à divers visuels symboliques de la ville de Paris.	24.5 %
Ambiguity	A tiny precision or imprecision in the question makes the plausible answer in the context incorrect.	Question: Quelle est la **dernière** station de la ligne ? Context: [...] La ligne se dirige vers l'est en position axiale jusqu'à **la station Balard** [...]	17.6 %
Out-of-context	While some concepts of the question are discussed in the context, at least one key concept of the question is not mentioned in the context.	Question: Quelle était **la profession** de Nicolas Bachelier ? Context: Les projets les plus réalistes sont présentés au roi au XVIe siècle. Un premier projet est présenté par Nicolas Bachelier en 1539 aux États de Languedoc, puis un second en 1598 par Pierre Reneau, et enfin un troisième projet proposé par Bernard Arribat de Béziers en 1617 [...]	6.9 %
Semantical Similarity	All concepts of the question are mentioned in the context, while the question remains unanswered in the context	Question: Quel est le nom du troisième volet de la saga ? Context: [...] Le fait que Solo soit plongé dans la carbonite constitue en outre une alternative pour les scénaristes si Harrison Ford refuse de jouer dans le troisième volet de la saga. En effet, George Lucas n'est pas assuré que sa vedette accepte de reprendre à nouveau le rôle après son succès dans **Les Aventuriers de l'arche perdue**.	29.4 %

Table 2: Categories of adversarial questions and their respective proportion in a FQUAD2.0 sample of 102 questions. Bold words are the plausible answers or discriminative terms within the question. Colored terms are co-references between question and context.

Model	Dataset	EM	F1	F1$_{has\ ans}$	NoAns$_{F1}$	NoAns$_P$	NoAns$_R$
CamemBERT$_{BASE}$	FQuAD2.0	63.3	68.7	82.5	62.1	82	49.9
CamemBERT$_{LARGE}$	FQuAD2.0	**78**	**83**	**90.1**	**82.3**	**93.6**	**73.4**

Table 3: Baseline results on the FQuAD2.0 validation set while training is made on the expanded training set containing 13,591 unanswerable questions.

brary (Wolf et al., 2019). All models were fine-tuned on 3 epochs, with a warmup ratio of 6%, a batch size of 16 and a learning rate of $1.5 \cdot 10^{-5}$. The optimizer used is AdamW with its default parameters. All experiments were carried out on a single Nvidia V100 16 GB GPU. Whenever necessary, gradient accumulation was used to train with batch size not fitting within the GPU memory. The results obtained on the FQuAD2.0 development set for the different metrics are presented in Table 3.

These first results allow us to draw the following conclusions:

- One can see that the best trained model, CamemBERT$_{LARGE}$, obtains a rather high score of 82.3 % for the NoAns$_{F1}$ metric, while keeping a high score of 90.1 % for the F1$_{has\ ans}$ metric. It confirms that it is possible for a pre-trained French Language Model to learn to determine with high precision when a French question is unanswerable, while extracting the correct answer in most cases when a question is answerable.

- As observed in (d'Hoffschmidt et al., 2020) or (Micheli et al., 2020), Question Answering seems to be a complex task for a small size (base, small)

Model	FQuAD1.1		FQuAD2.0	
	EM	F1	EM	F1
CamemBERT$_{BASE}$	78.1	88.1	73.1	82.5
CamemBERT$_{LARGE}$	82.4	91.8	81.3	90.1

Table 4: Comparison of scores obtained on the FQuAD1.1 dev set for models trained on FQuAD1.1 or FQuAD2.0.

fine-tuned Language Model to solve, and hence the obtained performances are highly dependent to model size, bigger models performing much better than smaller ones. It appears that for Adversarial Question Answering this observation is even more important, with CamemBERT$_{LARGE}$ scoring a 20.2 % absolute improvement in NoAns$_{F1}$ metric compared to CamemBERT$_{BASE}$.

5.2. Comparison with FQuAD1.1 scores

Whilst the models presented in the previous subsection clearly learned to both extract accurate answers from answerable questions and determine when a question is unanswerable, one may also be interested
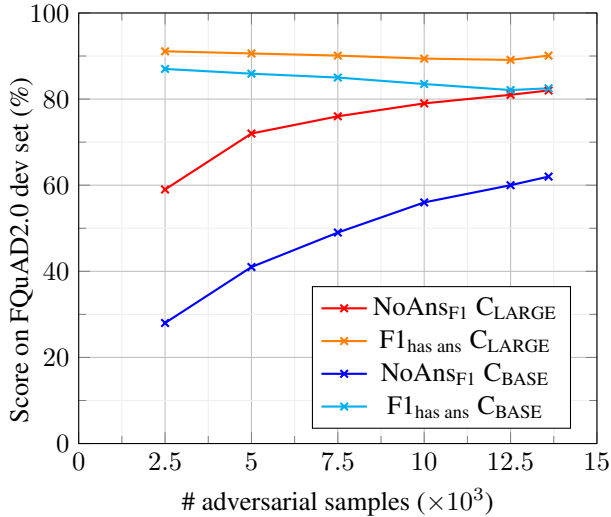
Figure 2: Evolution of $NoAns_{F1}$ and $F1_{has\ ans}$ for CamemBERT models depending on the number of unanswerable questions in the training dataset

in whether these models extract as accurate answers as similar models solely fine-tuned on FQuAD1.1, ie. only on answerable questions.

To answer this question, we present in Table 4 a comparison of our models of interest in two different set-ups: when fine-tuned solely on FQuAD1.1 and when fine-tuned on the entirety of FQuAD2.0. All evaluations are on FQuAD1.1 dev set. By dataset construction, the F1 score on the FQuAD1.1 dev set is strictly equivalent to the $F1_{has\ ans}$ on the FQuAD2.0 dev set. Results for fine-tuning on FQuAD1.1 are extracted from (d'Hoffschmidt et al., 2020).

With the addition of unanswerable questions during fine-tuning, the model is encouraged to predict that some questions are unanswerable. And as for every model, the $NoAns_P$ is strictly lower than 100%, there are answerable questions in the dev set, for which models tend to wrongly predict that they are unanswerable. Then for these questions, the predicted answer is the empty string instead of the expected answer. Hence, we can expect a decrease of the $F1_{has\ ans}$ metric in comparison to the set-up where a model is fine-tuned solely on FQuAD1.1.

This assumption is confirmed in Table 4 with a shrinking gap as model size grows. Indeed, $F1_{has\ ans}$ is only 1.7 absolute points lower for CamemBERT$_{LARGE}$ trained on FQuAD2.0 compaired to the same model fine-tuned solely on FQuAD1.1. For CamemBERT$_{BASE}$, the gap grows to 5.6 points. This gap evolution also follows the evolution of the $NoAns_P$ metric which is equal to 82% for CamemBERT$_{BASE}$ and 93.5% for CamemBERT$_{LARGE}$.

5.3. Learning curves

To get a better grasp of how many adversarial questions are needed for a model to learn to determine

when a question is unanswerable, we conduct several fine-tuning experiments with an increasing number of adversarial questions used for training. For every training, all answerable questions of the training set of FQuAD2.0 (i.e. the training set of FQuAD1.1), and unanswerable questions are progressively added to the training set with increments of 2500 questions. We conduct such experiments for the two model architectures of CamemBERT$_{BASE}$ and CamemBERT$_{LARGE}$. The results are displayed in Figure 2.

From these experiments, we observe the following:

- The CamemBERT$_{LARGE}$ model needs quite few adversarial examples before achieving decent performances. Indeed the model trained with 5k adversarial questions achieves 88% of the performance of the best model trained with 13.6k adversarial questions, which is 2.7 times more unanswerable questions.

- The slope of the CamemBERT$_{BASE}$ learning curve is higher than for CamemBERT$_{LARGE}$. For example, the CamemBERT$_{BASE}$ model trained with 5k adversarial questions achieves only 66% of the performance of the best CamemBERT$_{BASE}$ model trained with 13.6k adversarial questions. We conclude that the value brought by additional data is more important for smaller models than for bigger ones. However, we also observe that the CamemBERT$_{LARGE}$ model trained with 2.5k adversarial questions performs on par with the CamemBERT$_{BASE}$ model trained with 12.5k adversarial questions (5 times more data).

- Whatever the model, the learning curve has not flatten yet, which means that both architectures would benefit from more adversarial training samples. In order to do so, one would need to annotate further adversarial questions, which we leave for future work.

5.4. Baseline performances by question category

We present in section 3.3 a detailed analysis of the different challenges FQuAD2.0 adversarial questions provide. To understand how well the baseline CamemBERT models trained perform on each one of these challenges, we present in table 6 evaluation results on each of these categories. As the evaluation is solely made on adversarial questions, the chosen metric is the recall of the NoAns task: $NoAns_R$.

6. Multilingual experiments

The previous experiments focus on the study of fine-tuning French Language Models with the FQuAD2.0 dataset. However, one could ask the following questions:

- Could a multilingual Language Model fine-tuned solely on English Question Answering datasets

Model	Training Dataset	Test Dataset	EM	F1	F1$_{\text{has ans}}$	NoAns$_{\text{F1}}$
XLM-R$_{\text{BASE}}$	SQuAD2.0	FQuAD2.0	56	62.4	75.9	56.2
	SQuAD2.0 + FQuAD2.0	FQuAD2.0	**64.4**	**69.6**	78.4	**66.4**
CamemBERT$_{\text{BASE}}$	FQuAD2.0	FQuAD2.0	63.3	68.7	**82.5**	62.1
CamemBERT$_{\text{BASE}}$	FQuAD2.0*	FQuAD2.0*	60.5	66.1	83.5	56.4
RoBERTa$_{\text{BASE}}$	SQuAD2.0*	SQuAD2.0*	69.7	73.3	85.3	73.4
XLM-R$_{\text{LARGE}}$	SQuAD2.0	FQuAD2.0	67.3	73.4	87.8	68.1
	SQuAD2.0 + FQuAD2.0	FQuAD2.0	76.8	82.1	87.2	81.9
CamemBERT$_{\text{LARGE}}$	FQuAD2.0	FQuAD2.0	**78**	**83**	**90.1**	**82.3**

Table 5: Results for multilingual experiments with FQuAD2.0 and SQuAD2.0. Best score on FQuAD2.0 development set for both model sizes are highlighted in bold.

	NoAns$_{\text{R}}$	
Category	C$_{\text{BASE}}$	C$_{\text{LARGE}}$
Antonym	36	68
Entity Swap	36	86
Ambiguity	50	56
Out-of-context	43	71
Semantical Similarity	37	70

Table 6: Baseline models recalls for NoAns task on each category of adversarial questions. C$_{\text{BASE}}$ refers to CamemBERT$_{\text{BASE}}$.

compete against such a model, thanks to the existence of several large-scale English Question Answering datasets?

- Does the combination of French and English Question Answering datasets during training make a multilingual Language Model better than a monolingual Language Model in this Question Answering task?

We use FQuAD2.0 and SQuAD2.0, respectively as French and English reference datasets for these experiments. We benchmark two multilingual models: XLM-RoBERTa$_{\text{BASE}}$ and XLM-RoBERTa$_{\text{LARGE}}$ (Conneau et al., 2019) which are comparable in size respectively to CamemBERT$_{\text{BASE}}$ and CamemBERT$_{\text{LARGE}}$. Experimental set-up and parameters (relative to model sizes) are identical to the ones described in Section 5. Fine-tuning experiments are summarized in Table 5.
One can make the following observations from these results:

- Results in zero-shot setting are promising. We call zero-shot setting the FQuAD2.0 evaluations of models trained solely on SQuAD2.0 because the

models were not trained with any question-answer pair in French. For example, XLM-R$_{\text{LARGE}}$ reaches in zero-shot setting better performances on the FQuAD2.0 dataset than CamemBERT$_{\text{BASE}}$ trained on FQuAD2.0. Nevertheless, this observation must be put into perspective by reminding that SQuAD2.0 training set includes 43.5k adversarial questions, hence 3.2 times more than FQuAD2.0. By relying on the learning curves presented in Section 5, one can suppose that a CamemBERT$_{\text{BASE}}$ trained with 43.5k French adversarial questions would have substantially better performances than our actual best CamemBERT$_{\text{BASE}}$ model.

- For both model sizes, the CamemBERT model reaches better performances than the XLM-R model in the zero-shot setting with a substantial margin. One can then conclude than with 13.5k adversarial questions, we are beyond the point where training a French monolingual model on French question-answer pairs brings better results than using a multilingual model trained solely on English question-answer pairs.

- By combining FQuAD2.0 and SQuAD2.0 training sets, XLM-R$_{\text{BASE}}$ performs slightly better than CamemBERT$_{\text{BASE}}$ trained with FQuAD2.0, while for large models, CamemBERT is slightly better. We believe this is another demonstration of the ability of bigger language models to perform well in small data regimes, while smaller language models need large-scale datasets to reach their potential. Hence, it seems more interesting in a low computing resource setting and low training data availability setting to rely on multilingual models leveraging the more important availability of training data in English.

- As a matter of comparison with its English counterpart SQuAD2.0 and in order to assess the

relative difficulty of each dataset, we also performed experiments on subsets of SQuAD2.0 and FQuAD2.0, respectively denoted SQuAD2.0* and FQuAD2.0*, each with a training set of 50,000 answerable questions and 10,000 unanswerable ones. We finetuned a RoBERTa$_{BASE}$ model and a CamemBERT$_{BASE}$ with the same experimental set-ups for the Question Answering task. The scores obtained for all metrics are significantly better for the English set-up than the French one, notably with increases of 9.2% for the EM score and 6.2% for the overall F1 score. Besides, there is a 17% increase for the NoAns$_{F1}$ score for the English set-up. These results clearly indicate that the overall task for FQuAD2.0 is much harder than for SQuAD2.0 in terms of both the difficulty and the ambiguity of questions. In particular, the results for NoAns$_{F1}$ indicate that it is much harder for the French model to detect whether a question is answerable.

7. Conclusion & future work

In this paper, we introduced FQuAD2.0, a QA dataset with both answerable questions (coming from FQuAD1.1) and 17,000+ newly annotated unanswerable questions, for a total of almost 80,000 questions. To the best of our knowledge, this is the first French (and, perhaps most importantly, non-English) adversarial Question Answering dataset.

We trained various baseline models using Camem-BERT architectures. Our best model, a fine-tuned CamemBERT$_{LARGE}$, reaches 83% F1 score and 82.3% NoAns$_{F1}$, the latter measuring its ability to distinguish answerable questions from unanswerable ones. The study of learning curves with respect to the number of samples used for training such models show that our baseline models would benefit from additional unanswerable questions. In the future, we plan to collect additional samples to expand FQuAD2.0. For comparison, its english cousin SQuAD2.0 (Rajpurkar et al., 2018) contains 53,775 unanswerable questions. Such a large-scale dataset would of course enable the acquisition of even better models as the ones presented in Sections 5 and 6. As far as data collection is concerned, we could also collect additional answers for each unanswerable question. By following the same procedure as in (d'Hoffschmidt et al., 2020), this would allow for the computation of human performance, measuring the inherent difficulty of the challenge provided by FQuAD2.0.

On top of monolingual French experiments, we conducted various multilingual experiments demonstrating the relevance of using multilingual models fine-tuned on English resources to use in other languages when very few or no resources are available in this target language. Nevertheless, we also showed the superiority of a monolingual approach on the target language using a dataset such as FQuAD2.0 (as this was in our

case both economically and practically feasible). Besides, we also exhibited that FQuAD2.0 is probably a harder dataset of adversarial Question Answering than its English counterpart SQuAD2.0, as similar models perform better in the same conditions in English than in French.

Although the performances that we obtained on the FQuAD2.0 dataset are very good, the evaluation of the resulting models on other datasets is of crucial importance. In real-life industrial use cases, the contexts and questions asked vary from those present in FQuAD2.0: how can we perform efficient domain transfer on these datasets? This will be further evaluated in future iterations of this work.

Another topic of interest in the context of industrial use cases is the inference time of such large models. The best model we obtained is a CamemBERT$_{LARGE}$, but in some real-life applications where GPUs are unavailable or when we must handle a large number of requests in a short amount of time, we cannot afford the inference times that come with such large models. The use of smaller models than CamemBERT$_{LARGE}$ or CamemBERT$_{BASE}$ comes naturally into mind, such as LePetit (Micheli et al., 2020), also denoted as CamemBERT$_{SMALL}$. To overcome this limitation, we could use model compression techniques such as pruning (McCarley et al., 2019; Sanh et al., 2020), distillation (Hinton et al., 2015; Jiao et al., 2019; Sun et al., 2020) or quantization (Kim et al., 2021; Shen et al., 2019). We performed preliminary tests that seem very promising, they should be investigated further in the future.

9. Bibliographical References

Asai, A., Eriguchi, A., Hashimoto, K., and Tsuruoka, Y. (2018). Multilingual extractive reading comprehension by runtime machine translation. *CoRR*, abs/1809.03275.

Bahdanau, D., Cho, K., and Bengio, Y. (2016). Neural machine translation by jointly learning to align and translate.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Carrino, C. P., Costa-jussà, M. R., and Fonollosa, J. A. R. (2019a). Automatic spanish translation of the squad dataset for multilingual question answering. *ArXiv*, abs/1912.05200.

Carrino, C. P., Costa-jussà, M. R., and Fonollosa, J. A. R. (2019b). Automatic spanish translation of the squad dataset for multilingual question answering.

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W., Choi, Y., Liang, P., and Zettlemoyer, L. (2018). Quac : Question answering in context. *CoRR*, abs/1808.07036.

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., and Toutanova, K. (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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

d'Hoffschmidt, M., Belblidia, W., Heinrich, Q., Brendlé, T., and Vidal, M. (2020). FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online, November. Association for Computational Linguistics.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., and Gardner, M. (2019). Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. (2019). Tinybert: Distilling bert for natural language understanding.

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

Kabbadj, A. (2018). Something new in french text mining and information extraction (universal chatbot): Largest q&a french training dataset (110 000+).

Khashabi, D., Min, S., Khot, T., Sabharwal, A., Tafjord, O., Clark, P., and Hajishirzi, H. (2020). UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online, November. Association for Computational Linguistics.

Kim, S., Gholami, A., Yao, Z., Mahoney, M. W., and Keutzer, K. (2021). I-bert: Integer-only bert quantization.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. (2019). Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. H. (2017). RACE: large-scale reading comprehension dataset from examinations. *CoRR*, abs/1704.04683.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2019). Flaubert: Unsupervised language model pre-training for french.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., Villemonte de la Clergerie, É., Seddah, D., and Sagot, B. (2019). CamemBERT: a Tasty French Language Model. *arXiv e-prints*, page arXiv:1911.03894, Nov.

McCarley, J. S., Chakravarti, R., and Sil, A. (2019). Structured pruning of a bert-based question answering model.

Micheli, V., d'Hoffschmidt, M., and Fleuret, F. (2020). On the importance of pre-training data volume for compact language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858, Online, November. Association for Computational Linguistics.

Micheli, V., Heinrich, Q., Fleuret, F., and Belblidia, W. (2021). Structural analysis of an all-purpose question answering model.

Okazawa, S. (2021). Swedish translation of squad2.0. https://github.com/susumu2357/ SQuAD_v2_sv.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual bert? *CoRR*, abs/1906.01502.

Rachel, K., Guillaume, L., Mathilde, B., Frédéric, A., Gilles, M., Thomas, S., Edmundo-Pavel, S.-M., and Staiano, J. (2020). Project piaf: Building a native french question-answering dataset. In *Proceedings of the 12th Conference on Language Resources and Evaluation*. The International Conference on Language Resources and Evaluation.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (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, November. Association for Computational Linguistics.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Reddy, S., Chen, D., and Manning, C. D. (2018). Coqa: A conversational question answering challenge. *CoRR*, abs/1808.07042.

Saeidi, M., Bartolo, M., Lewis, P., Singh, S., Rocktäschel, T., Sheldon, M., Bouchard, G., and Riedel, S. (2018). Interpretation of natural language rules in conversational machine reading.

Sanh, V., Wolf, T., and Rush, A. M. (2020). Movement pruning: Adaptive sparsity by fine-tuning.

Shen, S., Dong, Z., Ye, J., Ma, L., Yao, Z., Gholami, A., Mahoney, M. W., and Keutzer, K. (2019). Q-bert: Hessian based ultra low precision quantization of bert. *CoRR*, abs/1909.05840.

Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., and Zhou, D. (2020). MobileBERT: a compact task-agnostic BERT for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170, Online, July. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zhu, M., Ahuja, A., Juan, D.-C., Wei, W., and Reddy, C. K. (2020). Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849, Online, November. Association for Computational Linguistics.