FQuAD: French Question Answering Dataset

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé
Illuin Technology
Paris, France
{martin, wacim, quentin, tom}@illuin.tech

Maxime Vidal
ETH Zurich
mvidal@student.ethz.ch

Abstract

Recent advances in the field of language modeling have improved state-of-the-art results on many Natural Language Processing tasks. Among them, Reading Comprehension has made significant progress over the past few years. However, most results are reported in English since labeled resources available in other languages, such as French, remain scarce. In the present work, we introduce the **F**rench **Qu**estion **A**nswering **D**ataset (FQuAD). FQuAD is a French Native Reading Comprehension dataset of questions and answers on a set of Wikipedia articles that consists of 25,000+ samples for the 1.0 version and 60,000+ samples for the 1.1 version. We train a baseline model which achieves an F1 score of 92.2 and an exact match ratio of 82.1 on the test set. In an effort to track the progress of French Question Answering models we propose a leaderboard and we have made the 1.0 version of our dataset freely available at https://illuin-tech.github.io/FQuAD-explorer/.

1 Introduction

Current progress in language modeling has led to increasingly successful results on various Natural Language Processing (NLP) tasks. This is namely the case of the Reading Comprehension task (Richardson et al., 2013). However, Reading Comprehension datasets are costly and difficult to collect and are essentially native English datasets. Indeed, datasets such as SQuAD1.1 (Rajpurkar et al., 2016), SQuAD2.0 (Rajpurkar et al., 2018), or CoQA (Reddy et al., 2018) have fostered important and impressive progress for English Question Answering models over the past few years. The lack of native language annotated datasets apart from English is one of the main reasons why the development of language specific Question Answering models is lagging behind and this is namely the case for French.

In order to fill the gap for the French language, we introduce a French Reading Comprehension dataset similar to SQuAD1.1. The dataset consists of French native questions and answers samples annotated by a team of university students. The dataset comes in two versions. First FQuAD1.0, containing over 25,000+ samples. Second, FQuAD1.1 containing over 60,000+ samples. The 35,000+ additional samples have been annotated with more demanding guidelines to strengthen complexity of the data and model to make the task harder. More specifically, the training, development, and test sets of FQuAD1.0 contain respectively 20,703, 3,188, and 2,189 samples. And the training, development, and test sets of FQuAD1.1 contain respectively 50,741, 5,668, and 5,594 samples.

In order to evaluate the FQuAD dataset, we perform various experiments by fine-tuning BERT based Question Answering models on both versions of the FQuAD dataset. The experiments involve the fine-tuning of French monolingual model CamemBERT (Martin et al., 2019), and multilingual models mBERT (Pires et al., 2019) and XLM-RoBERTa (Conneau et al., 2019).

We perform also two types of cross-lingual Reading Comprehension experiences. First, we evaluate the performance of the zero-shot cross-lingual transfer learning approach as stated in Artetxe et al. (2019) and Lewis et al. (2019) on our newly obtained native French dataset. Second, we evaluate the performance of the translation approach by fine-tuning models on the French translated version of SQuAD1.1. The results of these two experiments help to better understand how the two cross-lingual approaches actually perform on a native dataset.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1193–1208
November 16 - 20, 2020.

2 Related Work

The Reading Comprehension task (RC) (Richardson et al., 2013; Rajpurkar et al., 2016) attempts to solve the Question Answering (QA) problem by finding the text span in one or several documents or paragraphs that answers a given question (Ruder, 2020).

2.1 Reading Comprehension in English

Many Reading Comprehension datasets have been built in English. Among them SQuAD1.1 (Rajpurkar et al., 2016), then later SQuAD2.0 (Rajpurkar et al., 2018) has become one of the major reference dataset for training question answering models. Later, similar initiatives such as NewsQA (Trischler et al., 2016), CoQA (Reddy et al., 2018), QuAC (Choi et al., 2018), HotpotQA (Yang et al., 2018) have broadened the research area for English Question Answering.

These datasets are similar but each of them introduces its own subtleties. For instance, SQuAD2.0 (Rajpurkar et al., 2018) develops unanswerable adversarial questions. CoQA (Reddy et al., 2018) focuses on Conversation Question Answering in order to measure the ability of algorithms to understand a document and answer series of interconnected questions that appear in a conversation. QuAC (Choi et al., 2018) focuses on Question Answering in Context developed for Information Seeking Dialog (ISD). The benchmark established by Yatskar (2018) offers a qualitative comparison of these datasets. Finally, HotpotQA(Yang et al., 2018) attempts to extend the Reading Comprehension task to more complex reasoning by introducing multi-hop questions where the answer must be found among multiple documents.

2.2 Reading Comprehension in other languages

Native Reading Comprehension datasets other than English remain rare. Among them, some initiatives have been carried out in Chinese, Korean and Russian and all of them have been built in a similar way to SQuAD1.1. The SberQuAD dataset (Efimov et al., 2019) is a Russian native Reading Comprehension dataset and is made up of 50,000+ samples. The CMRC 2018 (Cui et al., 2019) dataset is a Chinese native Reading Comprehension dataset that gathers 20,000+ question and answer pairs. The KorQuAD dataset (Lim et al., 2019) is a Korean native Reading Comprehension dataset that is made up of 70,000+ samples. Note that following our work, the PIAF project (Rachel et al., 2020) has released a native French Dataset of 3,835 question and answer pairs. A complete overview of the aforementioned datasets is given as additional material in appendix A in table 8.

As language specific datasets are costly and challenging to obtain, an alternative consists in developing cross-lingual models that can transfer to a target language without requiring training data in that language (Lewis et al., 2019). It has indeed been shown that these unsupervised multilingual models generalize well in a zero-shot cross-lingual setting (Artetxe et al., 2019). For this reason, cross-lingual Question Answering has recently gained traction and two cross-lingual benchmarks have been released, i.e. XQuAD (Artetxe et al., 2019) and MLQA (Lewis et al., 2019). The XQuAD dataset (Artetxe et al., 2019) is obtained by translating 1,190 question and answer pairs from the SQuAD1.1 development set by professionals translators in 10 foreign languages. The MLQA dataset (Lewis et al., 2019) consists of over 12,000 question and answer samples in English and 5,000 samples in 6 other languages such as Arabic, German and Spanish. Note that the two aforementioned datasets do not cover French.

Another alternative consists in translating the training dataset into the target language and fine-tuning a language model on the translated dataset. This is namely the case of Carrino et al. (2019) where the authors develop a specific translation method called Translate Align Retrieve (TAR) to translate the English SQuAD1.1 dataset into Spanish. The resulting Spanish SQuAD1.1 dataset is used to fine-tune a multilingual model that reaches a performance of respectively 68.1/48.3% F1/EM and 77.6/61.8% F1/EM on MLQA cross-lingual benchmark (Lewis et al., 2019) and XQuAD (Artetxe et al., 2019). Note that a similar approach has been adopted for French and Japanese in Asai et al. (2018) and Siblini et al. (2019). In Siblini et al. (2019) a multilingual BERT is trained on English texts of SQuAD1.1, and evaluated on the small translated Asai et al. French corpus. This set-up reaches a promising score of 76.7/61.8 % F1/EM. Another translation approach was also explored in Kabbadj (2018) where the whole SQuAD1.1 dataset was translated and adapted to French with the Google Translate API.

2.3 Language modeling for Reading Comprehension

Increasingly efficient language models have been released recently such as GPT-2 (Radford et al., 2018), BERT (Devlin et al., 2018), XLNet (Yang et al., 2019) and RoBERTa (Liu et al., 2019). They have indeed disrupted the Reading Comprehension task and most of NLP fields: pre-training a language model on a generic corpus, eventually fine-tuning it on a domain specific corpus and then training it on a downstream task is the de facto state-of-the-art approach for optimizing both performances and annotated data volumes (Devlin et al., 2018; Liu et al., 2019). For instance, the top performing models on the SQuAD1.1 and SQuAD2.0 leaderboards[1] are essentially transformer based models. Unfortunately, the aforementioned models are pre-trained on English corpora and their use for French is therefore limited.

Multilingual models pre-trained on large multilingual datasets attempt to alleviate the language specific shortcoming characteristic of the former models such as Lample and Conneau (2019), Pires et al. (2019) and more recently XLM-R (Conneau et al., 2019). It has been shown in Conneau et al. (2019), Artetxe et al. (2019) and Lewis et al. (2019) that multilingual models are flexible and perform reasonably well on other languages than English. However, they do not appear to perform better than specific language models (Lewis et al., 2019).

Regarding French, few resources were available until recently. First, the CamemBERT models (Martin et al., 2019) were trained on 138 GB of French text from the Oscar dataset (Ortiz Suárez et al., 2019). Second, the FlauBERT models (Le et al., 2019) were trained on 71 GB of text. Note that both models were pre-trained with the Masked Language Modeling task only (Martin et al., 2019; Le et al., 2019). Both models reach similar performances on French NLP tasks such as PoS, NER and NLI. However, their performance has not yet been evaluated on the Reading Comprehension task as no French dataset is available.

3 Dataset Collection

The collection was conducted in two distinct steps: the first one resulted in FQuAD1.0 with 25,000+ question and answer pairs, and the second one resulted in FQuAD1.1 with 60,000+ question and answer pairs. Apart from that, the collection follows the same standards and guidelines as SQuAD1.1 (Rajpurkar et al., 2016).

3.1 Paragraphs collection

A set of 1,769 articles are collected from the French Wikipedia page referencing quality articles [2]. From this set, a total of 145 articles are randomly sampled to build the FQuAD1.0 dataset. Also, 181 additional articles are randomly sampled to extend the dataset to FQuAD1.1. resulting in a total of 326 articles. Among them, articles are randomly assigned to the training, development, and test sets. The training, development, and test sets for FQuAD1.0 are respectively made up of 117, 18, and 10 articles. For the FQuAD1.1 dataset, they are respectively made up of 271, 30, and 25 articles. Note that train, development, test split is performed at the article level in order to avoid any possible biases.

The paragraphs that are at least 500 characters long are kept for each article, similarly to Rajpurkar et al. (2016). This technique results in 4,951, 768, and 523 paragraphs for respectively the training, development, and test sets of FQuAD1.0. For FQuAD1.1, the number of collected paragraphs for the same sets are respectively 12,123, 1,387, and 1,398.

3.2 Question and answer pairs collection

A specific annotation platform was developed to collect the question and answer pairs. The workers are French students that were hired in collaboration with the Junior Enterprise of CentraleSupélec [3]. They were paid about 16.5 euros per hour of work. The guidelines for writing question and answer pairs for each paragraph are the same as for SQuAD1.1 (Rajpurkar et al., 2016). First, the paragraph is presented to the student on the platform and the student reads it. Second, the student thinks of a question whose answer is a span of text within the context. Third, the student selects the smallest span in the paragraph which contains the answer. The process is then repeated until 3 to 5 questions are generated and correctly answered. The students were asked to spend on average 1 minute on each question and answer pair. This amounts to an average of 3-5 minutes per annotated paragraph. Additionally during the annotation process, about 25 % of the questions for each annotator were manually reviewed to make sure the questions remain

[1] rajpurkar.github.io/SQuAD-explorer

[2] https://fr.wikipedia.org/wiki/Categorie:Article_de_qualite

[3] https://juniorcs.fr/en/

of high quality. Final dataset metrics are shared in table 2.

3.3 Additional answers collection

Additional answers are collected to decrease the annotation bias similarly to Rajpurkar et al. (2016). For each question in the development and test sets, two additional answers are collected, resulting in three answers per question for these sets. The crowd-workers were asked to spend on average 30 seconds to answer each question.

For the same question, several answers may be correct: for instance the question *Quand fut couronné Napoléon ?* would have several possible answers such as *mai 1804*, *en mai 1804*, or *1804*. As all those answers are admissible, enriching the test set with several annotations for the same question, with different annotators, is a way to decrease annotation bias. The additional answers are useful to get an indication of the human performance on FQuAD.

3.4 FQuAD1.0 & FQuAD 1.1

The results for the first annotation process resulting in the FQuAD1.0 dataset are reported in table 1. The number of collected question and answer pairs amounts to 26,108. Diverse analysis to measure the difficulty of the resulting dataset are performed as described in the next section. A complete annotated paragraph is displayed in figure 2.

Dataset	Articles	Paragraphs	Questions
Train	117	4,921	20,731
Development	18	768	3,188
Test	10	532	2,189

Table 1: The number of articles, paragraphs and questions for FQuAD1.0

The first dataset is extended with additional annotation samples to build the FQuAD1.1 dataset reported in table 2. The total number of questions amounts to 62,003. The FQuAD1.1 training, development and test sets are then respectively composed of 271 articles (83%), 30 (9%), and 25 (8%). Following the version 1.0 annotation campaign, we observed that the most difficult questions for the models trained were questions of types *Why* and *How* or answers involving `verbs` and `adjectives`. This is further explained in section E. Therefore, we asked the annotators to come up with more questions of these specific types. The motivation was to come up with more challenging questions to understand if the trained models could improve on those. This constitutes the only difference with the first annotation process. The additional answer collection process remains the same.

Dataset	Articles	Paragraphs	Questions
Train	271	12,123	50,741
Development	30	1,387	5,668
Test	25	1,398	5,594

Table 2: The number of articles, paragraphs and questions for FQuAD1.1

4 Dataset Analysis

4.1 Answer analysis

To analyse the collected answers, a combination of rule-based regular expressions and entity extraction using spaCy (Honnibal and Montani, 2017) are used. First, a set of regular expression rules are applied to isolate `dates` and other `numerical` answers. Second, `person` and `location` entities are extracted using Named Entity Recognition. Third, a rule based approach is adopted to extract the remaining `proper nouns`. Finally, the remaining answers are labeled into `common noun`, `verb`, and `adjective` phrases, or `other` if no labels were found. Answer type distribution is shown in table 3.

Answer type	Freq [%]	Example
Common noun	26.6	rencontres
Person	14.6	John More
Other proper nouns	13.8	Grand Prix d'Italie
Other numeric	13.6	1,65 m
Location	14.1	Normandie
Date	7.3	1815
Verb	6.6	être dépoussiéré
Adjective	2.6	méprisant, distant et sec
Other	0.9	gimmick

Table 3: Answer type by frequency for the development set of FQuAD1.1

4.2 Question analysis

The second analysis aims at understanding the question types of the dataset. The present analysis is performed rule-based only. Table 4 first demonstrates that the annotation process issued a wide range of question types, underlining the fact that *What (que)* represents almost half (47.8%) of the corpus. This important proportion may be explained by this

formulation encompassing both the English *What* and *Which*, as well as a possible natural bias in the annotators way of asking questions. Our intuition is that this bias is the same during inference, as it originates from native French structure.

Question	Freq [%]	Example
What (que)	47.8	Quel pays parvient à ...
Who	12.2	Qui va se marier bientôt ?
Where	9.6	Où est l'échantillon ...
When	7.6	Quand a eu lieu la ...
Why	5.3	Pourquoi l'assimile ...
How	6.8	Comment est le prix ...
How many	5.6	Combien d'albums ...
What (quoi)	4.1	De quoi est faite la ...
Other	1	Donner un avantage de ...

Table 4: Question type by frequency for the development set of FQuAD1.1

4.3 Question-answer differences

The difficulty in finding the answer given a particular question lies in the linguistic variation between the two. This can come in different ways, which are listed in table 9 The categories are taken from Rajpurkar et al. (2016): *Synonymy* implies key question words are changed to a synonym in the context; *World knowledge* implies key question words require world knowledge to find the correspondence in the context; *Syntactic variation* implies a difference in the structure between the question and the answer; *Multiple sentence reasoning* implies knowledge requirement from multiple sentences in order to answer the question. We randomly sampled 6 questions from each article in the development set and manually labeled them. Note that samples can belong to multiple categories.

4.4 Evaluation metrics

The Exact Match (EM) and F1-score metrics are common metrics being computed to evaluate the performances of a model. The former measures the percentage of predictions matching exactly one of the ground truth answers. The later computes the average overlap between the predicted tokens and the ground truth answer. The prediction and ground truth are processed as bags of tokens. For questions labeled with multiple answers, the F1 score is the maximum F1 over all the ground truth answers.

The evaluation process in Rajpurkar et al. (2016) for both the F1 and EM ignores some English punctuation, i.e. the *a*, *an*, *the* articles. In order to remain consistent with the former approach, the French evaluation process ignores the following articles: *le, la, les, l', du, des, au, aux, un, une*.

4.5 Human performance

Similarly to SQuAD, human performances are evaluated on the development and test sets in order to assess how humans agree on answering questions. This score gives a comparison baseline when assessing the performance of a model. To measure the human performance, for each question, two of the three answers are considered as the ground truth, and the third as the prediction. In order not to bias this choice, the three answers are successively considered as the prediction, so that three human scores are calculated. The three runs are then averaged to obtain the final human performance for the F1 Score and Exact Match. For the test set and development set we find a Human Score reaching respectively 91.2% F1 and 75.9% EM, and 91.2% F1 and 78.3% EM. An in-depth analysis is carried out in appendix C to compare the FQuAD1.1 to SQuAD1.1 in terms of Human Performance and answer length.

5 Experiments

5.1 Experimental set-up

The experimental set-up is kept the same across all the experiments. The number of epochs is set to 3, with a learning rate equal to $3.0 \cdot 10^{-5}$. The learning rate is scheduled according to a warm-up linear scheduler where the percentage ratio for the warm-up is consistently set to 6%. The batch size is kept constant across the training and is equal to 8 for the base models and 4 for the large ones. The optimizer that is being used is `AdamW` with its default parameters. All the experiments were carried out with the HuggingFace transformers library (Wolf et al., 2019) on a single V100 GPU.

5.2 Native French Reading Comprehension

The goal of these experiments is two fold. First, we want to evaluate the performance of the French language models $\text{CamemBERT}_{\text{BASE}}$ and $\text{CamemBERT}_{\text{LARGE}}$ (Martin et al., 2019) on FQuAD. Second, we want to evaluate the performances of multilingual models using the same set-up. For this purpose we train two multilingual models, i.e. mBERT (Pires et al., 2019) and the XLM-RoBERTa models (Conneau et al., 2019). Finally, we compare the results for both the monolingual and multilingual models to understand how they

Reasoning	Example	Frequency
Synonymy	Question: Quel est le sujet **principal** du film ? Context: Le sujet **majeur** du film est le *conflit de Rick Blaine entre l'amour et la vertu* : il doit choisir entre...	35.2 %
World knowledge	Question: Quand John Gould a-t-il décrit la nouvelle **espèce d'oiseau** ? Context: **E. c. albipennis** décrite par John Gould en *1841*, se rencontre dans le nord du Queensland, l'ouest du golfe de Carpentarie dans le Territoire du Nord et dans le nord de l'Australie-Occidentale.	11.1 %
Syntactic variation	Question: Combien d'**auteurs ont parlé de la merveille du monde de Babylone** ? Context: Dès les premières campagnes de fouilles, on chercha **la « merveille du monde » de Babylone : les Jardins suspendus décrits par** *cinq* auteurs...	57.4 %
Multiple sentence reasoning	Question: Qu'est ce qui rend la situation de menace des cobs précaire ? Context: En 1982, les chercheurs en concluent que **le cob normand** est victime de consanguinité, de dérive génétique et de la disparition de ses structures de coordination. *L'âge avancé de ses éleveurs* rend **sa** situation précaire.	17.6 %

Table 5: Question-answer relationships in 108 randomly selected samples from the FQuAD development set. In bold the elements needed for the corresponding reasoning, in italics the selected answer.

perform on the French dataset. Note that for each experiment, the fine-tuning is performed on the training set of FQuAD1.1 and evaluated on the development and test sets of FQuAD.1.1. Additional fine-tuning experiments performed on the training set of FQuAD1.0 are presented in appendix D.

5.3 Cross-lingual Reading Comprehension

Cross-lingual Reading comprehension follows mainly two approaches as explained in section 2. On one hand, experiments carried out in Lewis et al. (2019) and Artetxe et al. (2019) evaluate how multilingual models fine-tuned on the English SQuAD1.1 dataset perform on other languages such as Spanish, Chinese or Arabic. On the other hand, initiatives such as Carrino et al. (2019) attempt to translate the dataset in the target language to fine-tune a model. The newly obtained FQuAD dataset makes it now possible to test both approaches on the English-French cross-lingual setup. Note however that French is unfortunately not supported by the cross-lingual benchmark proposed by Lewis et al. (2019); Artetxe et al. (2019).

First, we perform several experiments with a so called zero-shot learning approach. In other words, we fine-tune multilingual models on the English SQuAD1.1 dataset and we evaluate them on the FQuAD1.1 development set. In addition to that, the opposite approach is also carried out, i.e. fine-tuned models on FQuAD1.1 are evaluated on the SQuAD1.1 development set.

Second, we fine-tune CamemBERT on the SQuAD1.1 training dataset translated into French. For this purpose, the SQuAD1.1 training set is translated using NMT (Ott et al., 2018). Note that the translation process makes it difficult to keep all the samples from the original dataset and, for the sake of simplicity, we discard the translated answers that do not align with the start/end positions of the translated paragraphs. The resulting translated dataset *SQuAD1.1-fr-train* contains about 40,700 question and answer pairs. The fine-tuned model is then evaluated on the native French FQuAD1.1 development set.

6 Results

6.1 Native French Reading Comprehension

The training experiments on FQuAD1.1-train are summed up in table 6. Note that experiments carried out on FQuAD1.0-train are available in the appendix in table 12. All the models are evaluated on the FQuAD1.1 test and development sets.

	FQuAD1.1-test		FQuAD1.1-dev	
Model	F1	EM	F1	EM
Human Perf.	91.2	75.9	92.1	78.3
CamemBERT$_{BASE}$	88.4	78.4	88.1	78.1
CamemBERT$_{LARGE}$	**92.2**	**82.1**	**91.8**	**82.4**
mBERT	86.0	75.4	86.2	75.5
XLM-R$_{BASE}$	85.9	75.3	85.5	74.9
XLM-R$_{LARGE}$	89.5	79.0	89.1	78.9

Table 6: Results of the experiments for various monolingual and multilingual models carried out on the training dataset of **FQuAD1.1-train** and evaluated on test and development sets of FQuAD1.1

Monolingual models The CamemBERT$_{BASE}$ trained on FQuAD1.1 reaches 88.4% F1 and 78.4% EM as reported on 6. Interestingly, the base version surpasses the Human Score in terms of Exact Match on the test set. The best model, CamemBERT$_{LARGE}$ trained on FQuAD1.1 reaches a performance of 92.2% F1 and 82.1% EM on the test set, which is the highest score across the experiments and surpasses already the Human Performance for both metrics on the test and development sets. By means of comparison, the best model of the SQuAD1.1 leaderboard reaches 95.1% F1 and 89.9% EM on the SQuAD1.1 test set (Yang et al., 2019). Note that while the size of FQuAD1.1 remains smaller than its english counterpart, the aforementioned results yield a very promising baseline. Note further that the same model reaches a performance of 93.3% F1 and 84.6% EM on the test set of FQuAD1.0, hereby supporting the fact that FQuAD1.1 includes more difficult question.

Multilingual models The results of the experiments carried out for the multilingual models reported in 6 show that they perform also very well when evaluated on the test and development sets of FQuAD1.1. The top performer in this category is XLM-R$_{LARGE}$ which reaches 89.5% F1 and 79% EM on FQuAD1.1-test. The model XLM-R$_{BASE}$ scores 85.9% F1 and 75.3% EM on the test set. Comparatively, mBERT model reaches a similar performance with 86.0% F1 and 75.4% EM. These experiments show that monolingual language models reach stronger performances than multilingual models overall. Nevertheless, it is important to note that XLM-R$_{LARGE}$ model performs better than CamemBERT$_{BASE}$ on both the test and development sets and even surpasses the Human Performance in terms of Exact Match on the test set.

6.2 Cross-lingual Reading Comprehension

The results for the experiments on the cross-lingual set-up are reported in table 7. On one hand, the French monolingual models are fine-tuned on the French translated version of SQuAD1.1 and evaluated on the development set of FQuAD1.1. On the other hand, multi-language models are fine-tuned respectively on SQuAD1.1 and FQuAD1.1 and then evaluated respectively on the development sets of FQuAD1.1 and SQuAD1.1 in order to evaluate the performance of zero-shot learning set-up.

Translated Reading Comprehension First, the results for CamemBERT$_{BASE}$ fine-tuned on the French translated version of SQuAD1.1. show a performance of 81.8% F1 and 67.8% EM as reported in 7. Compared to CamemBERT$_{BASE}$ fine-tuned on FQuAD, this result is about 6.3 points less effective in terms of F1 score and even more important in terms of EM score, i.e. 10.3. Second, the results for CamemBERT$_{LARGE}$ show an improved performance of 87.5% F1 and 73.9% EM. Compared to the native version, this result is lower by 4.3 points in terms of F1 Score and 8.5 points in terms of EM.

Even if the translated dataset contains about 40,700 question and answer pairs, while the train set of FQuAD1.1 contains 50,700 pairs, such a difference does not find roots in varying datasets sizes as another lead experiment whose results are described in section E demonstrated that training a CamemBERT$_{BASE}$ model on 40,000 question and answer pairs results in only a 0.4 absolute point difference regarding F1-score as opposed to training on 50,000 question and answer pairs.

These experiments show therefore that models fine-tuned on translated data do not perform as well as when they are fine-tuned on native dataset. This difference is probably explained by the fact that NMT produces translation inaccuracies that impact the EM score more than F1 score. When we merge the native and the translated dataset into what we call the Augmented dataset, we do not observe a significant performance improvement. Interestingly, the CamemBERT$_{LARGE}$ model performs slightly worse when fine-tuned on translated samples.

Zero-shot learning To evaluate how multi-language models transfer on other languages similarly to Lewis et al. (2019) and Artetxe et al. (2019), we report the results of our experiments with XLM-R$_{BASE}$ and XLM-R$_{LARGE}$ in 7. We find that XLM-R$_{BASE}$ trained on FQuAD1.1 reaches 83.0% F1 and 73.5 % EM on the SQuAD1.1 dev set. When trained on SQuAD1.1 it reaches 81.4% F1 and 68.4% EM on the FQuAD1.1 dev set. Next, we find that XLM-R$_{LARGE}$ reaches 88.8% F1 and 79.5% on the SQuAD1.1 dev set when trained on FQuAD1.1 and 86.1% F1 and 73.2% EM on the FQuAD1.1 dev set when trained on SQuAD1.1. The results show that the models perform very well compared to the results when trained on the native French and native English datasets. Indeed, XLM-R$_{BASE}$ shows a drop of only 4.1% and 6.5% in terms of F1 and EM score on the FQuAD1.1 dev set when

Model	Train Dataset	SQuAD1.1-dev F1 [%]	SQuAD1.1-dev EM [%]	FQuAD1.1-dev F1 [%]	FQuAD1.1-dev EM [%]
Human Perf.		91	80.5	92.1	78.3
$\text{CamemBERT}_{\text{BASE}}$	FQuAD1.1	-	-	88.1	78.1
	SQuAD1.1-fr	-	-	81.8	67.8
	Augmented	-	-	88.3	78.0
$\text{CamemBERT}_{\text{LARGE}}$	FQuAD1.1	-	-	91.8	82.4
	SQuAD1.1-fr	-	-	87.5	73.9
	Augmented	-	-	91.2	81.6
$\text{XLM-R}_{\text{BASE}}$	FQuAD1.1	83.0	73.5	85.5	74.9
	SQuAD1.1	88.1	80.9	81.4	68.4
$\text{XLM-R}_{\text{LARGE}}$	FQuAD1.1	88.8	79.5	89.1	78.9
	SQuAD1.1	90.7	83.4	86.1	73.2

Table 7: Results for the zero-shot learning experiments on the SQuAD1.1 and FQuAD1.1 development sets

compared to the model trained on the native french samples. And $\text{XLM-R}_{\text{LARGE}}$ show a drop on 3.0% and 5.7% in terms of F1 and EM score. Note that the same relationship can be observed for the model trained on FQuAD1.1 and evaluated on SQuAD1.1 although the drop in performance is slightly less important. Interestingly, the large models perform in general very well on the cross-lingual zero-shot set-up.

7 Discussion

7.1 Monolingual vs. multilingual language models

Through our language models benchmark on FQuAD, we have evaluated several monolingual and multilingual models. The $\text{CamemBERT}_{\text{BASE}}$ and $\text{CamemBERT}_{\text{LARGE}}$ models reach a very promising baseline and the large model even outperforms the Human Performance consistently across the development and test datasets.

For comparable model sizes we find that the monolingual models outperform multilingual models on the Reading Comprehension task. However, we find that multilingual models such as mBERT (Pires et al., 2019) or $\text{XLM-R}_{\text{BASE}}$ and $\text{XLM-R}_{\text{LARGE}}$ (Conneau et al., 2019) reach very promising scores. We find that $\text{XLM-R}_{\text{LARGE}}$ performs consistently better than the monolingual model $\text{CamemBERT}_{\text{BASE}}$ on both the development and test sets of FQuAD1.1. Let us further highlight that $\text{XLM-R}_{\text{LARGE}}$ reaches 79% EM on FQuAD-test which is better than Human Performance, while the F1 score remains only 2% below it. As such a model is pre-trained on a multilingual corpus, we can hope that it could be used with reasonable performances on other languages.

7.2 Translated Reading Comprehension

Fine-tuning $\text{CamemBERT}_{\text{BASE}}$ on a French translated dataset yields 81.8/67.8% F1/EM on the FQuAD1.1 dev set. By means of comparison, $\text{CamemBERT}_{\text{BASE}}$ scores 88.1/78.1% F1/EM on the same set when trained with native French data. We find here that there exists an important gap between both approaches. Indeed, models that are fine-tuning on native data outperform models fine-tuned on translated data by an order of magnitude of 10% for the Exact Match.

In Carrino et al. (2019), the authors report a performance of 77.6/61.8% F1/EM score when mBERT is trained on a Spanish-translated SQuAD1.1 and evaluated on XQuAD (Artetxe et al., 2019). While the two approaches differ in terms of evaluation dataset, i.e. XQuAD is not a native Spanish dataset, and model, mBERT vs. CamemBERT, and although French and Spanish are different languages, they are close enough in their construction and structure, so that comparing these two approaches is relevant to us. Given the level of effort put into the translation process in Carrino et al. (2019), we think that both translation-based approaches, although using very recent language models, reach a performance ceiling with translated data. We observe also that enriching native French training data with the translated samples does not improve the performances on the native evaluation set. Given our experiments, we conclude therefore that there exist a significant gap between the native French and the French translated data in terms on quality and indicates that approaches based on translated data reach ceiling

performances.

7.3 Cross-lingual Reading Comprehension

The zero-shot experiments show that multilingual models can reach strong performances on the Reading Comprehension task in French or English when the model has not encountered labels of the target language. For example, the XLM-R_{LARGE} model fine-tuned solely on FQuAD1.1 reaches a performance on SQuAD just a few points below the English Human Performance. The same is also observed while fine-tuning solely on SQuAD1.1 and evaluating on the development set of FQuAD1.1. We conclude here in agreement with Artetxe et al. (2019) and Lewis et al. (2019) that the transfer of models from French to English and vice versa relevant approach when no annotated samples are available in the target language.

The experiments also show that the zero-shot performances are better for SQuAD than for FQuAD. This phenomenon can be explained by structural differences between French and English or an increased difficulty of FQuAD compared to SQuAD. It is also possible that the XLM-R language models used are capturing English language specifics better than for other languages because the dataset used for pre-training these models contains more English data. Further experiments aiming at training multilingual models on both FQuAD1.1 and SQuAD1.1 may improve the results further. This possibility is left for future works.

8 Conclusion

In the present work, we introduce the **F**rench **Qu**estion **A**nswering **D**ataset. The contexts are collected from the set of high quality Wikipedia articles. With the help of French college students, 60,000+ questions have been manually annotated. The FQuAD dataset is the result of two different annotation processes. First, FQuAD1.0 is collected to build a 25,000+ questions dataset. Second, the dataset is enriched to reach 60,000+ questions resulting in FQuAD1.1. The development and test sets have both been enriched with additional answers for the evaluation process.

We find that the Human performances for FQuAD1.1 on the test and development sets reach respectively a F1-score of 91.2% and an Exact Match of 75.9%, and a F1-score of 92.1% and an Exact Match of 78.3%. Furthermore, we find that the Human performances on FQuAD1.1 reach comparable scores to SQuAD1.1.

Various experiments were carried out to evaluate the performances of monolingual and multilingual language models. Our best model, $CamemBERT_{LARGE}$, achieves a F1-score and an Exact Match of respectively 92.2% and 82.1%, surpassing the established Human performance in terms of F1-Score and Exact Match. The experiments show that multilingual models reach promising results but monolingual models of comparable sizes perform better.

The FQuAD1.0 training and FQuAD1.1 development sets are made publicly available in order to foster research in the French NLP area. We believe our dataset can boost French research in other NLP fields such as NLU, Information Retrieval or Open Domain Question Answering to cite a few. The extension of the dataset to adversarial questions similarly to SQuAD2.0 is left for future works.

Acknowledgments

We would like to warmly thank Robert Vesoul, Co-Director of CentraleSupélec's Digital Innovation Chair and CEO of Illuin Technology, for his help and support in enabling and funding this project while leading it through.

We would also like to thank Enguerran Henniart, Lead Product Manager of Illuin annotation platform, for his assistance and technical support during the annotation campaign.

Finally we extend our thanks to the whole Illuin Technology team for their reviewing and constructive feedbacks.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *ArXiv*, abs/1910.11856.

Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *CoRR*, abs/1809.03275.

Casimiro Pio Carrino, Marta Ruiz Costa-jussà, and José A. R. Fonollosa. 2019. Automatic spanish translation of the squad dataset for multilingual question answering. *ArXiv*, abs/1912.05200.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac : Question answering in context. *CoRR*, abs/1808.07036.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5883–5889, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Pavel Efimov, Leonid Boytsov, and Pavel Braslavski. 2019. Sberquad - russian reading comprehension dataset: Description and analysis.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Ali Kabbadj. 2018. Something new in french text mining and information extraction (universal chatbot): Largest qa french training dataset (110 000+).

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2019. Flaubert: Unsupervised language model pre-training for french.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *ArXiv*, abs/1910.07475.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. Korquad1.0: Korean qa dataset for machine reading comprehension.

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.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. CamemBERT: a Tasty French Language Model. *arXiv e-prints*, page arXiv:1911.03894.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. *CoRR*, abs/1806.00187.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *CoRR*, abs/1906.01502.

Keraron Rachel, Lancrenon Guillaume, Bras Mathilde, Allary Frédéric, Moyse Gilles, Scialom Thomas, Soriano-Morales Edmundo-Pavel, and Jacopo Staiano. 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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners. *CoRR*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. Coqa: A conversational question answering challenge. *CoRR*, abs/1808.07042.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Sebastian Ruder. 2020. Nlp progress.

Wissam Siblini, Charlotte Pasqual, Axel Lavielle, and Cyril Cauchois. 2019. Multilingual question answering from formatted text applied to conversational agents. *ArXiv*, abs/1910.04659.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *CoRR*, abs/1611.09830.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mark Yatskar. 2018. A qualitative comparison of coqa, squad 2.0 and quac. *CoRR*, abs/1809.10735.

A Additional tables and figures

Table 8 lists some of the available Reading Comprehension datasets along with the number of samples they contain[4]. By means of comparison, Table 8 also includes FQuAD. Figure 2 is a screenshot of the annotation interface used to collect FQuAD. Last, figure 2 shows examples of question and answer pairs for a paragraph in FQuAD.

Dataset	Language	Size
SQuAD1.1	English	100,000+
SQuAD2.0	English	150,000+
NewsQA	English	100,000+
CoQA	English	127,000+
QuAC	English	98,000+
HotpotQA	English	113,000+
KorQuAD	Korean	70,000+
SberQuAD	Russian	50,000+
CMR-2018	Chinese	20,000+
FQuAD1.0	French	**25,000+**
FQuAD1.1	French	**60,000+**
PIAF	French	3,835

Table 8: Benchmark of existing Reading Comprehension datasets, including FQuAD.

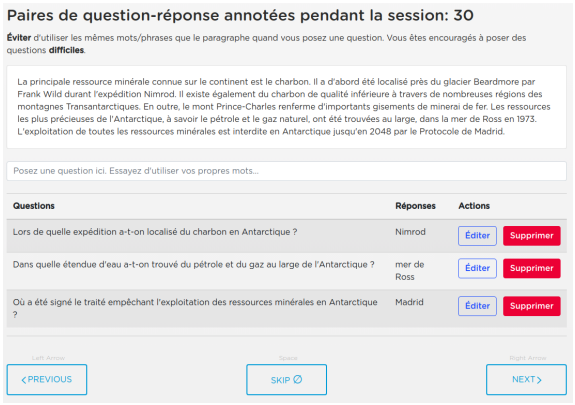

Figure 1: The interface used to collect the question/answers encourages workers to write difficult questions.

B Additional dataset analysis

B.1 Questions and answers differences

The difficulty in finding the answer given a particular question lies in the linguistic variation between the two. This can come in different ways, which are listed in table 9 The categories are taken from Rajpurkar et al. (2016): *Synonymy* implies key question words are changed to a synonym in the context; *World knowledge* implies key question words require world knowledge to find the correspondence in the context; *Syntactic variation* implies a difference in the structure between the question and the answer; *Multiple sentence reasoning* implies knowledge requirement from multiple sentences in order to answer the question. We randomly sampled 6 questions from each article in the development set and manually labeled them. Note that samples can belong to multiple categories.

[4] https://nlpprogress.com/english/question_answering.html

Article: Cérès
Paragraph:
Des observations de 2015 par la sonde Dawn ont confirmé qu'elle possède une forme sphérique, à la différence des corps plus petits qui ont une forme irrégulière. Sa surface est probablement composée d'un mélange de glace d'eau et de divers minéraux hydratés (notamment des carbonates et de l'argile), et de la matière organique a été décelée. Il semble que Cérès possède un noyau rocheux et un manteau de glace. Elle pourrait héberger un océan d'eau liquide, ce qui en fait une piste pour la recherche de vie extraterrestre. Cérès est entourée d'une atmosphère ténue contenant de la vapeur d'eau, dont deux geysers, ce qui a été confirmé le 22 janvier 2014 par l'observatoire spatial Herschel de l'Agence spatiale européenne.

Question 1: A quand remonte les observations faites par la sonde Dawn ?
Answer: *2015*

Question 2: Qu'ont montré les observations faites en 2015 ?
Answer: *elle possède une forme sphérique, à la différence des corps plus petits qui ont une forme irrégulière*

Question 3: Quelle caractéristique possède Cérès qui rendrait la vie extraterrestre possible ?
Answer: *un océan d'eau liquide*

Figure 2: Question answer pairs for a sample passage in FQuAD

B.2 The accrued difficulty of FQuAD1.1 vs FQuAD1.0

The table 10 reports the Human performances obtained for FQuAD1.0 and FQuAD1.1. The human score on FQuAD1.0 reaches 92.1% F1 and 78.4% EM on the test set and 92.6% and 79.5% on the development set. On FQuAD1.1, it reaches 91.2% F1 and 75.9% EM on the test set and 92.1% and 78.3% on the development set. We observe that there is a noticeable gap between the human performance on FQuAD1.0 test dataset and the human performance on the new samples of FQuAD1.1 with 78.4% EM score on the 2,189 questions of FQuAD1.0 test set and 74.1% EM score on the 3,405 new ques-

Reasoning	Example	Frequency
Synonymy	Question: Quel est le sujet **principal** du film ? Context: Le sujet **majeur** du film est le *conflit de Rick Blaine entre l'amour et la vertu* : il doit choisir entre...	35.2 %
World knowledge	Question: Quand John Gould a-t-il décrit la nouvelle **espèce d'oiseau** ? Context: **E. c. albipennis** décrite par John Gould en *1841*, se rencontre dans le nord du Queensland, l'ouest du golfe de Carpentarie dans le Territoire du Nord et dans le nord de l'Australie-Occidentale.	11.1 %
Syntactic variation	Question: Combien d'**auteurs ont parlé de la merveille du monde de Babylone** ? Context: Dès les premières campagnes de fouilles, on chercha **la « merveille du monde » de Babylone : les Jardins suspendus décrits par** *cinq* auteurs...	57.4 %
Multiple sentence reasoning	Question: Qu'est ce qui rend la situation de menace des cobs précaire ? Context: En 1982, les chercheurs en concluent que **le cob normand** est victime de consanguinité, de dérive génétique et de la disparition de ses structures de coordination. *L'âge avancé de ses éleveurs* rend **sa** situation précaire.	17.6 %

Table 9: Question-answer relationships in 108 randomly selected samples from the FQuAD development set. In bold the elements needed for the corresponding reasoning, in italics the selected answer.

tions of FQuAD1.1 test set. As explained in section 3 we insisted in our annotation guidelines of FQuAD1.1 that the questions should be more difficult. This gap in human performance constitutes for us a proof that answering to FQuAD1.1 new questions is globally more difficult than answering to FQuAD1.0 questions, hence making the final FQuAD1.1 dataset even more challenging.

Dataset	F1 [%]	EM [%]
FQuAD1.0-test.	92.1	78.4
FQuAD1.1-test	91.2	75.9
"FQuAD1.1-test new samples"	90.5	74.1
FQuAD1.0-dev	92.6	79.5
FQuAD1.1-dev	92.1	78.3
"FQuAD1.1-dev new samples"	91.4	76.7

Table 10: Human Performance on FQuAD

C Comparing FQuAD1.1 and SQuAD1.1

The SQuAD1.1 dataset (Rajpurkar et al., 2016) reports a human score for the test set equal to 91.2% F1 and 82.3% EM. Comparing the English score with the French ones, we notice that they are the same in terms of F1 score but differ by 6% on the Exact Match. This difference indicates a potential structural difference between FQuAD1.1 and SQuAD1.1. To better understand it we first compare the answer type distributions, then we compare the answer lengths for both datasets and finally we explore how the evaluation score varies with the answer length.

Answer type distribution The comparison in answer type distribution between the FQuAD1.1 and SQuAD1.1 datasets are reported in table 11. For both datasets, the most represented answer type is `Common Noun` with FQuAD1.1 scoring 26.6% and SQuAD1.1 scoring 31.8%. The less represented ones are `Adjective` and `Other` which have a noticeable higher proportion for SQuAD1.1 than FQuAD1.1 Compared to SQuAD1.1, a significant difference exists on structured entities such as `Person`, `Location`, and `Other Numeric` where FQuAD1.1 consistently scores above SQuAD1.1 with the exception of the `Date` category where FQuAD scores less. Based on these observations, it is difficult to understand the difference in human score between the two datasets.

Answer type	FQuAD1.1 [%]	SQuAD1.1 [%]
Common noun	26.6	31.8
Person	14.6	12.9
Other proper nouns	13.8	15.3
Location	14.1	4.4
Date	7.3	8.9
Other numeric	13.6	10.9
Verb	6.6	5.5
Adjective	2.6	3.9
Other	0.9	2.7

Table 11: Answer type comparison for the development sets of FQuAD1.1 and SQuAD1.1

Answer length To compare the answer lengths for the FQuAD1.1 and SQuAD1.1 datasets, we first remove every punctuation signs as well as

respectively french words *le, la, les, l', du, des, au, aux, un, une* and english words *a, an, the*. Then answers are split on white spaces to compute the number of tokens for each answer. The results are reported in figure 3. It appears clearly that FQuAD answers are generally longer than SQuAD answers. Furthermore, to highlight this important difference it is interesting to realise that the average number of tokens per answer for SQuAD1.1 is equal to 2.72 while it is equal to 4.24 for FQuAD1.1. This indicates that reaching a high Exact Match score on FQuAD is more difficult than on SQuAD.

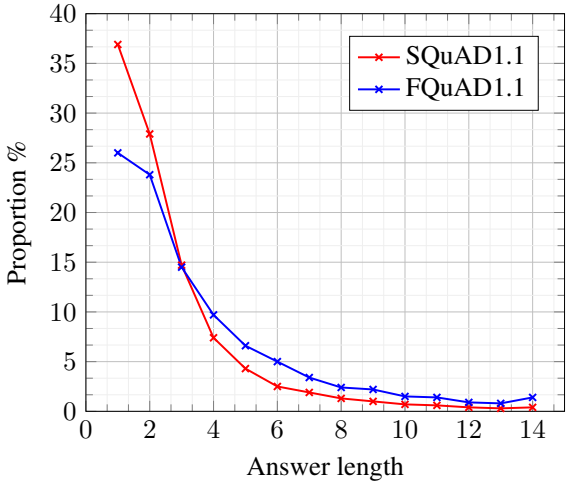

Figure 3: Answers lengths distribution for FQuAD and SQuAD

Human performance as a function of the answer length To understand if the answer length can impact the difficulty of the Reading Comprehension task, we group question and answer pairs in FQuAD and SQuAD by the number of tokens for each answer. The figure 4 shows the human performance as a function of the answer length. On one hand, it is straightforward to notice that the Exact Match quickly declines with an increasing answer length for both FQuAD and SQuAD. On the other hand, the F1 score is a lot less affected by answer length for both datasets. We conclude from these distributions that the difference in answers lengths between FQuAD and SQuAD may explain part of the difference in human performance regarding EM metric, while it does not seem to have an impact on human performance regarding F1 metric. And indeed, human performance regarding F1 metric is very similar between FQuAD and SQuAD. It is possible that these variations in answers lengths distributions are due to structural differences between French and English languages.

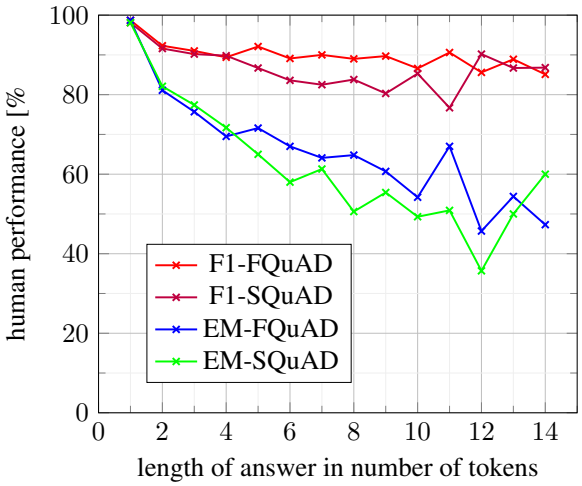

Figure 4: Evolution of the F1 and EM human scores for the answers length of the development sets of FQuAD1.1 and SQuAD1.1

Number of answers per question As indicated in Rajpurkar et al. (2018), the SQuAD1.1 and SQuAD2.0 development and test sets have on average 4.8 answers per question. By means of comparison, the FQuAD1.1 datasets has on average 3 answers per question for the development and test sets. The more answers to a question there are, the more likely it is that any other answer is equal to one of the expected answers. As a consequence, the higher number of answers in SQuAD1.1 contributes to the higher human performance compared to FQuAD1.1 regarding the exact match metric.

D Additional experiments

Training on FQuAD1.0 As we open source the 1.0 version of FQuAD dataset, we also reproduce all the native French Reading Comprehension fine-tuning experiments described in section 5.2 with the training set of FQuAD1.0.

Performance analysis An analysis of the predictions for the best trained model on FQuAD is carried out. We have explored the distribution of answer and questions types in section 4 and we report now the performance of the model in terms of F1 score and Exact Match for each category. This analysis aims at understanding how the model performs on the various question and answer types.

Learning curve The question of how much data is needed to train a question answering model remains relatively unexplored. In our effort of an-

notating FQuAD1.0 and FQuAD1.1 we have consistently monitored the scores to know if the annotation process must be continued or stopped. For this purpose, we present a learning curve obtained on the FQuAD1.1 test set by training CamemBERT$_{\text{BASE}}$ on an increasing number of question and answer samples. Both the EM and F1 scores are reported on the learning curve.

PIAF The French Dataset PIAF has been released after the first release of the present work. In order to assess the impact of the PIAF released samples (3,885 training samples), we perform two experiments using PIAF. First, we evaluate the CamemBERT models fine-tuned on FQuAD1.0 on the new samples. Second, we concatenate FQuAD1.0 and PIAF to train a new model and evaluate them on the test set of FQuAD1.1 to understand if the new samples bring additional score.

E Additional results

Training on FQuAD1.0 The experiments results are reported in table 12.

	FQuAD1.1-test		FQuAD1.1-dev	
Model	F1	EM	F1	EM
Human Perf.	91.2	75.9	92.1	78.3
CamemBERT$_{\text{BASE}}$	86.0	75.8	85.5	74.1
CamemBERT$_{\text{LARGE}}$	**91.5**	**82.0**	**91.0**	**81.2**
mBERT	83.9	72.3	83.1	71.8
XLM-R$_{\text{BASE}}$	82.2	71.4	82.4	71.0
XLM-R$_{\text{LARGE}}$	88.7	78.5	88.2	77.5

Table 12: Results of the experiments for various monolingual and multilingual models carried out on the training dataset of **FQuAD1.0-train** and evaluated on test and development sets of FQuAD1.1

Performance analysis Our best model CamemBERT$_{\text{LARGE}}$ is used to run the performance analysis on the question and answer types. Tables 13 and 14 present the results sorted by F1 score. The model performs very well on structured data such as `Date`, `Numeric`, or `Location`. Similarly, the model performs well on questions seeking for structured information, such as `How many`, `Where`, `When`. The `Person` answer type human score is very high on EM metric, meaning that these answers are easier to detect exactly probably because the answer is in general short. On the other end, the `How` and `Why` questions that probably expect a long and wordy answer are among the least well addressed. Note that `Verb` answers EM score is also quite low. This is probably due to either the variety of forms a verb can take, or to the fact that verbs are often part of long and wordy answers, which are by definition difficult to match exactly. Some prediction examples are available in the appendix. Selected samples are not part of FQuAD, but were sourced from Wikipedia.

Question Type	$F1$	EM	$F1_h$	EM_h
How many	**96.3**	**87.8**	93.3	82.1
When	96.1	83.3	92.6	78.3
Who	93.1	87.7	95.7	90.5
Where	92.7	74.3	88.4	66.5
What (que)	91.8	76.6	91.3	77.6
Why	91.5	61.9	88.1	56.8
What (quoi)	89.8	64.9	88.3	66.1
How	88.5	70.5	88.4	70.1
Other	77.8	53.3	84.7	58.3

Table 13: Performance on question types. $F1_h$ and EM_h refer to human scores

Answer Type	$F1$	EM	$F1_h$	EM_h
Date	**95.8**	**82.1**	92.6	78.1
Other	94.6	75.6	84.4	63.7
Location	92.8	80.7	92.0	78.5
Other numeric	92.8	79.1	91.7	76.7
Person	92.5	80.8	93.4	82.6
Other proper nouns	92.5	78.3	91.9	78.0
Common noun	91.3	74.4	89.8	73.1
Adjective	89.6	73.1	90.8	71.6
Verb	88.5	58.7	87.7	60.9

Table 14: Performance on answer types. $F1_h$ and EM_h refer to human scores

Learning curve The learning curve is obtained by performing several experiments with an increasing number of question and answer samples randomly taken from the FQuAD1.1 dataset. For each experiment, CamemBERT$_{\text{BASE}}$ is fine-tuned on the training subset and is evaluated on the FQuAD1.1 test set. The F1 scores and Exact Match are reported on the figure 5 with respect to the number of samples involved in the training. The figure shows that both the F1 and EM score follow the same trend. First, the model is quickly improving upon the first 10,000 samples. Then, F1 and EM are progressively flattening upon augmenting the number of training samples. Finally, they reach a

maximum value of respectively 88.4% and 78.4%. The results show us that a relatively low number of samples are needed to reach acceptable results on the reading comprehension task. However, to outperform the Human Score, i.e. 91.2% and 75.9 %, a larger number of samples is required. In the present case CamemBERT$_{\text{BASE}}$ outperforms the Human Exact Match after it is trained on 30,000 samples or more.

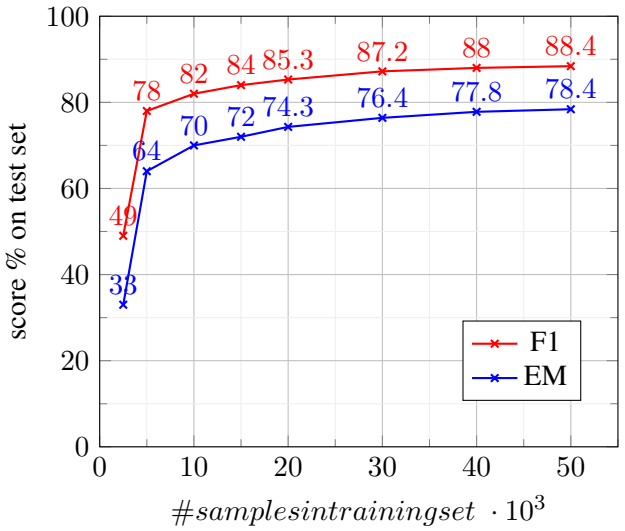

Figure 5: Evolution of the F1 and EM scores for CamemBERT$_{\text{BASE}}$ depending on the number of samples in the training dataset

PIAF Dataset The experiments carried out on PIAF are reported in table 15. To ease the comparison we also add the results from table 12. The results show that the F1 and EM performances reach a significantly lower level than on FQuAD1.1-test. One of the reasons for such a gap is the fact that the PIAF dataset does not include several answers per question as it is the case in SQuAD1.1 or in the present work.

	PIAF		FQuAD1.1-test	
Training data	F1	EM	F1	EM
FQuAD1.0 *(1)*	68.15	48.79	86.0	75.8
FQuAD1.0 *(2)*	74.43	54.39	91.5	82.0
FQuAD1.0 + PIAF *(1)*	-	-	86.8	76.2

Table 15: Results of the experiments for CamemBERT trained on **FQuAD1.0-train** and evaluated on PIAF. *(1)* has been trained with CamemBERT$_{\text{BASE}}$, *(2)* has been trained with CamemBERT$_{\text{LARGE}}$.