LREC 2022

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

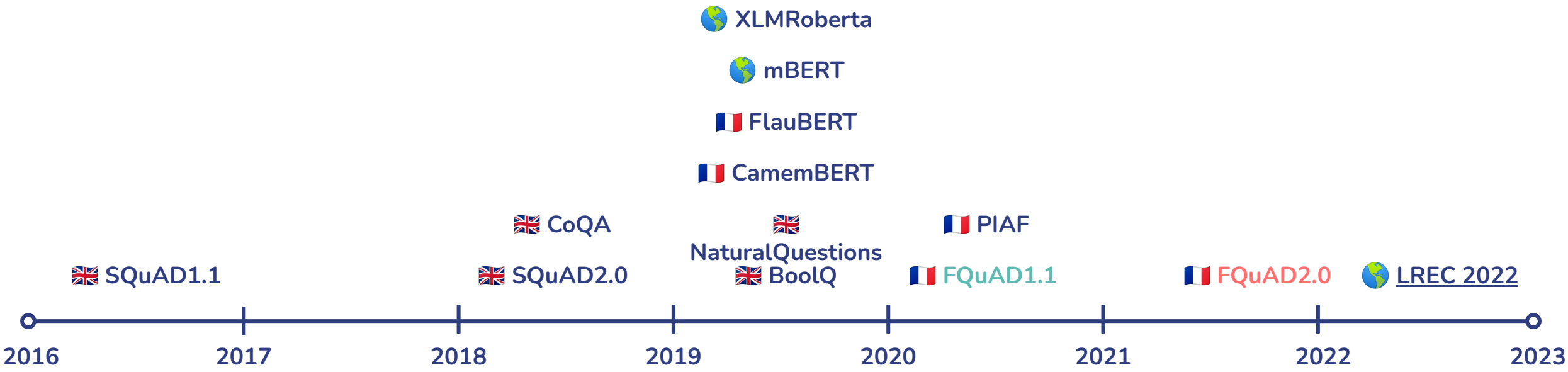
Quentin HEINRICH Gautier VIAUD Wacim BELBLIDIA



20 juin 2022

ILLUIN TECHNOLOGY

Introduction



Limitations of FQuAD1.1

- A model was 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 questions do not have an answer in the associated context

Introduction

Extractive Question - With Answer

Context

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.

Question

À quand remontent les observations faites par **Dawn** ?

Expected Answer

2015

Adversarial Question - Without Answer

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.

Question

Quel est le nom du **troisième volet de la saga** ?

Expected Answer

NO_ANSWER

Our contributions

- We introduce the FQuAD2.0 dataset, which extends FQuAD1.1 with **17,000+ unanswerable questions**, hand-crafted to be difficult to distinguish from answerable questions, making it the first non-English 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. 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 finetuned solely on question-answer pairs of a single language (English) perform in another language (French)**. We also take interest in performances of such models trained on both French and English datasets.

Dataset Collection & Analysis

Annotation guidelines in the adversarial framework

- 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

18 French annotators **7** mns / par. **4** adv. questions / par. **3100** paragraphs **17765** adv. questions

	FQuAD1.1			FQuAD2.0		
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**
Articles	271	30	25	271	30	25
Paragraphs	12123	1387	1398	12123	1387	1398
Answerable questions	50741	5668	5594	50741	5668	5594
Unanswerable questions	0	0	0	9481	4174	4110
Total questions	50741	5668	5594	60222	9842	9704

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

Dataset Collection & Analysis

Annotation guidelines in the adversarial framework

- 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

18 French annotators **7** mns / par. **4** adv. questions / par. **3100** paragraphs **17765** adv. questions

	FQuAD1.1			FQuAD2.0		
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**
Articles	271	30	25	271	30	25
Paragraphs	12123	1387	1398	12123	1387	1398
Answerable questions	50741	5668	5594	50741	5668	5594
Unanswerable questions	0	0	0	9481	4174	4110
Total questions	50741	5668	5594	60222	9842	9704

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

Dataset Collection & Analysis

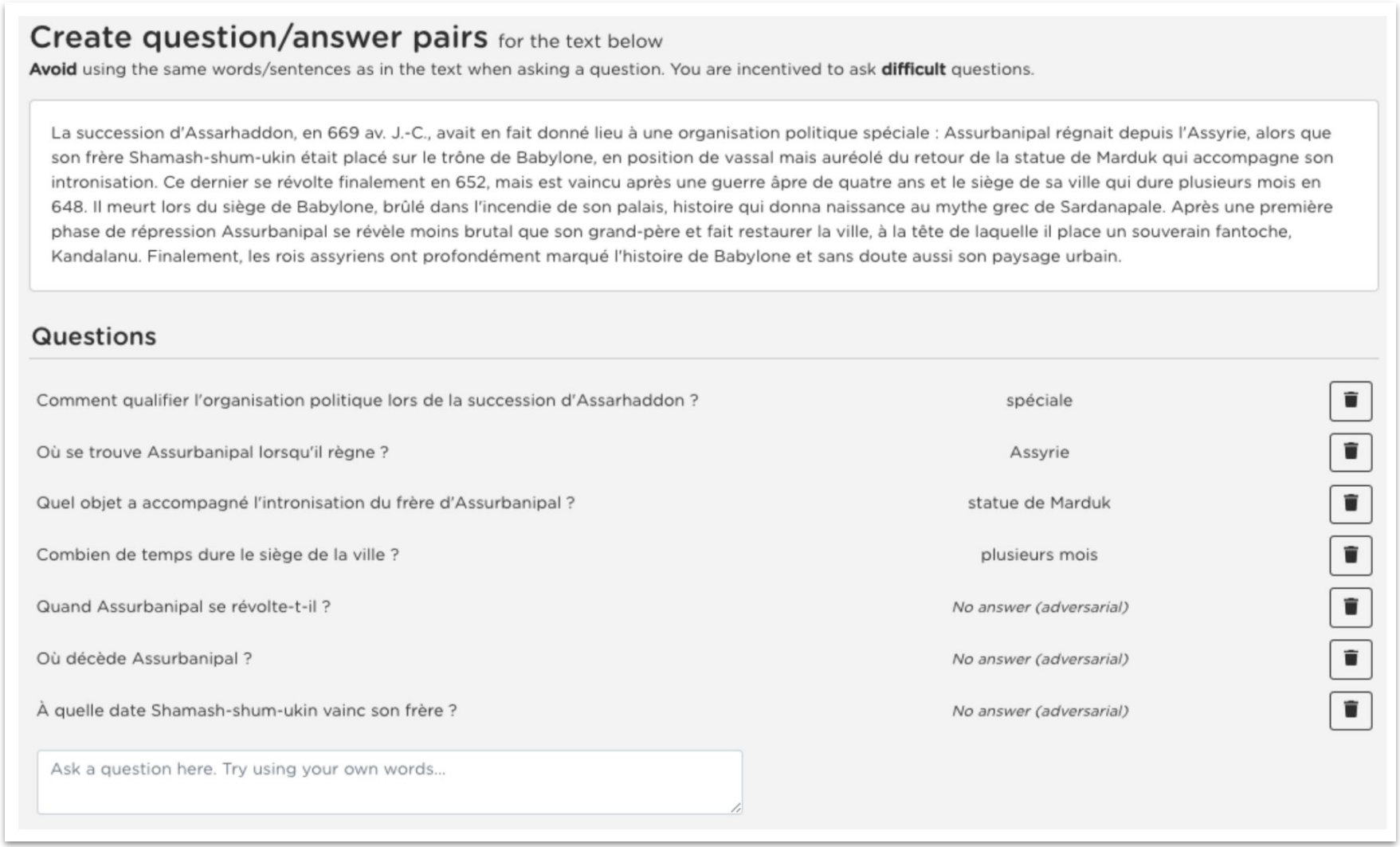


Figure 1: The interface used to collect the question-answer pairs for FQuAD. During the annotation process for FQuAD2.0, an annotator can see a paragraph and the associated answerable questions that were already collected for FQuAD1.1.

Dataset Collection & Analysis

Reasoning	Description	Example	Frequency
Antonym	Use of negation or antonym to make the question adversarial.	Question: Quels mammifères **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 Etats 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%
Semantic 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 coreferences between question and context.

Evaluation Metrics

Standard metrics

- **Exact Match (EM):** Percentage of predictions matching exactly one of the ground truth answers
- **F1-score (F1):** Average overlap between the predicted tokens and the ground truth answer

Remark: To extend these metrics, unanswerable questions are simply considered as answerable questions with the ground truth answer being an empty string

Extended metrics for Adversarial Question Answering

- $\mathbf{F1}_{\text{has ans}}$: Average F1 score, question-wise, as defined above, but limited to answerable questions
- $\mathbf{NoAns}_{\text{F1}}$: 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

Remarks

- $\mathbf{NoAns}_{\text{F1}}$ is a metric computed as a whole on the entirety of the FQuAD2.0 development set as a classification problem, while $\mathbf{F1}_{\text{has ans}}$ is computed question-wise and is an average of the individual scores for each question
- The global **F1-score** is **not** the weighted average of $\mathbf{F1}_{\text{has ans}}$ and $\mathbf{NoAns}_{\text{F1}}$

French Monolingual Experiments

Experiments

- Finetuning of models on the Question Answering task as in (Devlin et al., 2018) for:
  - **CamemBERT$_{LARGE}$** (24 layers, 1024 hidden dimensions, 12 attention heads, 340M parameters)
  - **CamemBERT$_{BASE}$** (12 layers, 768 hidden dimensions, 12 attention heads, 110M parameters)

3 epochs **6** % warmup ratio **16** batch size **$1.5 \cdot 10^{-5}$** learning rate | **V100 16 GB GPU**

Model	Dataset	EM	F1	F1$_{\text{has ans}}$	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.

French Monolingual Experiments

Experiments

- Finetuning of models on the Question Answering task as in (Devlin et al., 2018) for:
  - **CamemBERT$_{LARGE}$** (24 layers, 1024 hidden dimensions, 12 attention heads, 340M parameters)
  - **CamemBERT$_{BASE}$** (12 layers, 768 hidden dimensions, 12 attention heads, 110M parameters)

3 epochs **6** % warmup ratio **16** batch size **$1.5 \cdot 10^{-5}$** learning rate | **V100 16 GB GPU**

Model	Dataset	EM	F1	F1$_{\text{has ans}}$	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**

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.

French Monolingual Experiments

Experiments

- Finetuning of models on the Question Answering task as in (Devlin et al., 2018) for:
  - **CamemBERT$_{LARGE}$** (24 layers, 1024 hidden dimensions, 12 attention heads, 340M parameters),
  - **CamemBERT$_{BASE}$** (12 layers, 768 hidden dimensions, 12 attention heads, 110M parameters).

3 epochs **6** % warmup ratio **16** batch size **$1.5 \cdot 10^{-5}$** learning rate | **V100 16 GB GPU**

Model	Dataset	EM	F1	F1$_{\text{has ans}}$	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**

Question Answering seems to be a complex task for small size models, even more so in the adversarial framework: CamemBERT$_{LARGE}$ scores a 20.2% absolute improvement in the NoAns$_{F1}$ metric compared to CamemBERT$_{BASE}$.

French Monolingual Experiments

How does the extractive performance of FQuAD2.0 within the adversarial framework compare to the full extractive FQUAD1.1?

- With the addition of unanswerable questions during fine-tuning, the model is encouraged to predict that some questions are unanswerable. With $NoAns_P < 100\%$, **there are answerable questions in the dev set for which models tend to wrongly predict that they are unanswerable**.
- 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.

	EM		F1		ΔF1
Model	**FQuAD1.1**	**FQuAD2.0**	**FQuAD1.1**	**FQuAD2.0**	
$CamemBERT_{BASE}$	78.1	73.1	88.1	82.5	**-5.6**
$CamemBERT_{LARGE}$	82.4	81.3	91.8	90.1	**-1.7**

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

French Monolingual Experiments

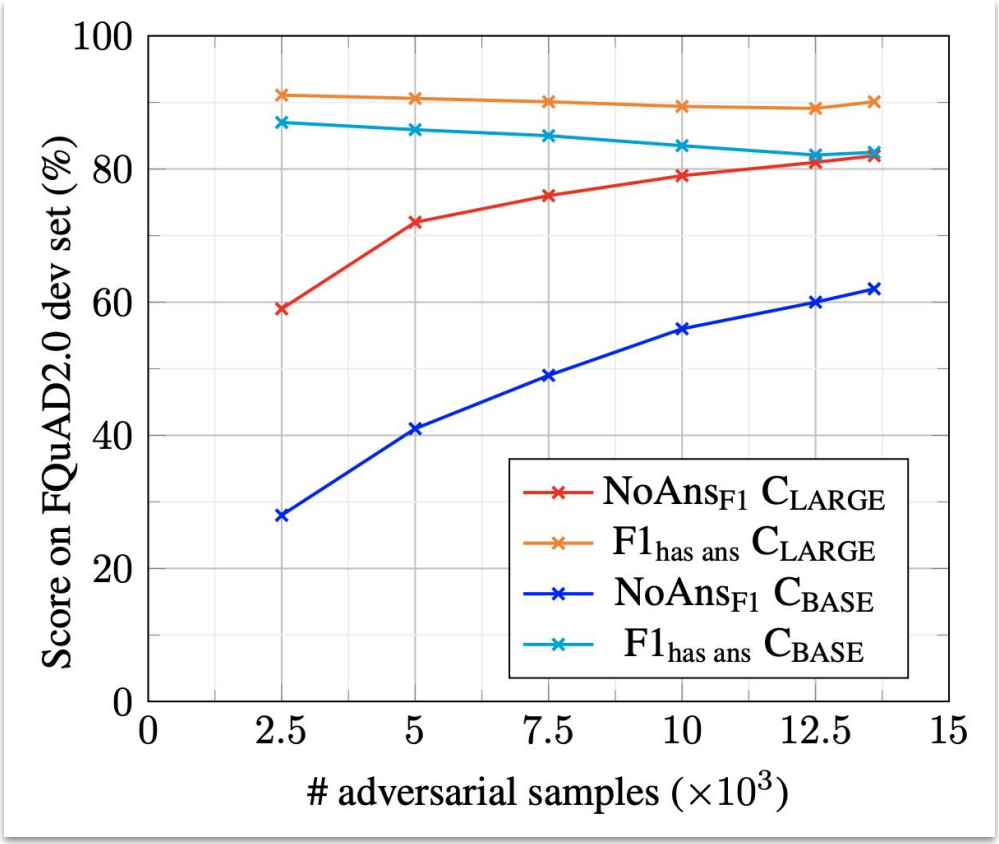


Figure 2: Evolution of NoAns_{F1} and $\text{F1}_{\text{has ans}}$ for CamemBERT models depending on the number of unanswerable questions in the training dataset.

How many adversarial questions are needed for a model to learn to determine when a question is unanswerable?

➤ 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 are used and unanswerable questions are progressively added with increments of 2500 questions.

French Monolingual Experiments

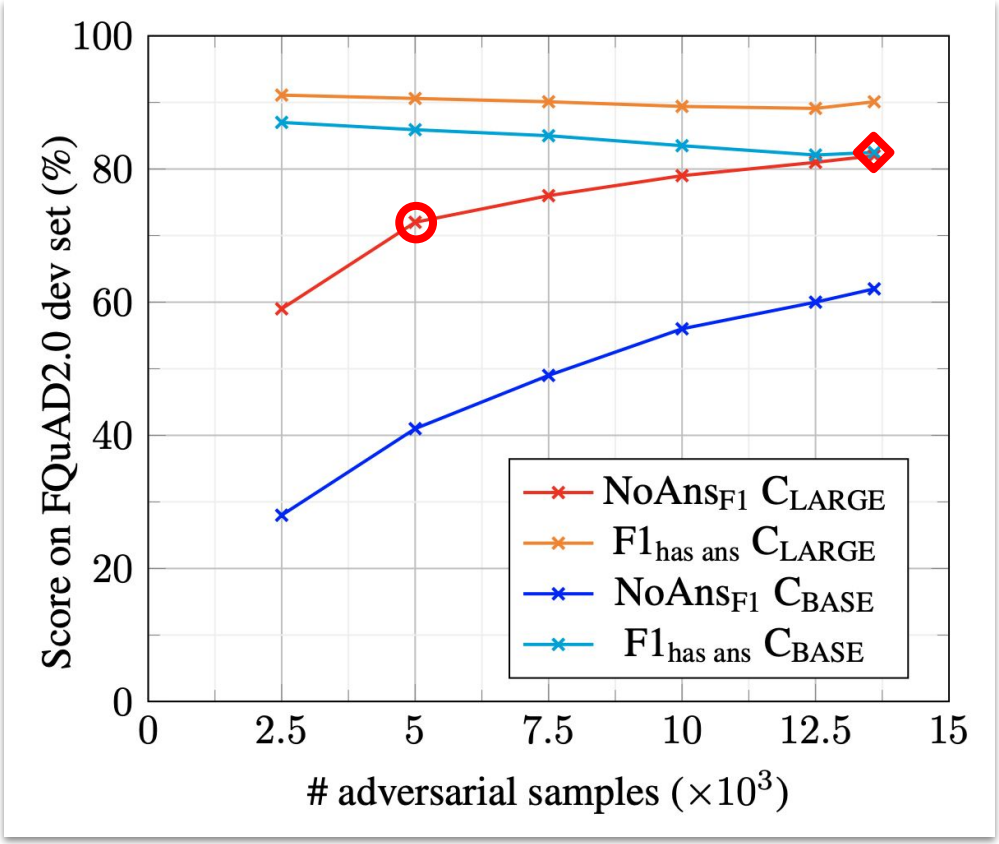


Figure 2: Evolution of $NoAns_{F1}$ and $F1_{has\ ans}$ for CamemBERT models depending on the number of unanswerable questions in the training dataset.

How many adversarial questions are needed for a model to learn to determine when a question is unanswerable?

➤ 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 are used and unanswerable questions are progressively added with increments of 2500 questions.

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.

French Monolingual Experiments

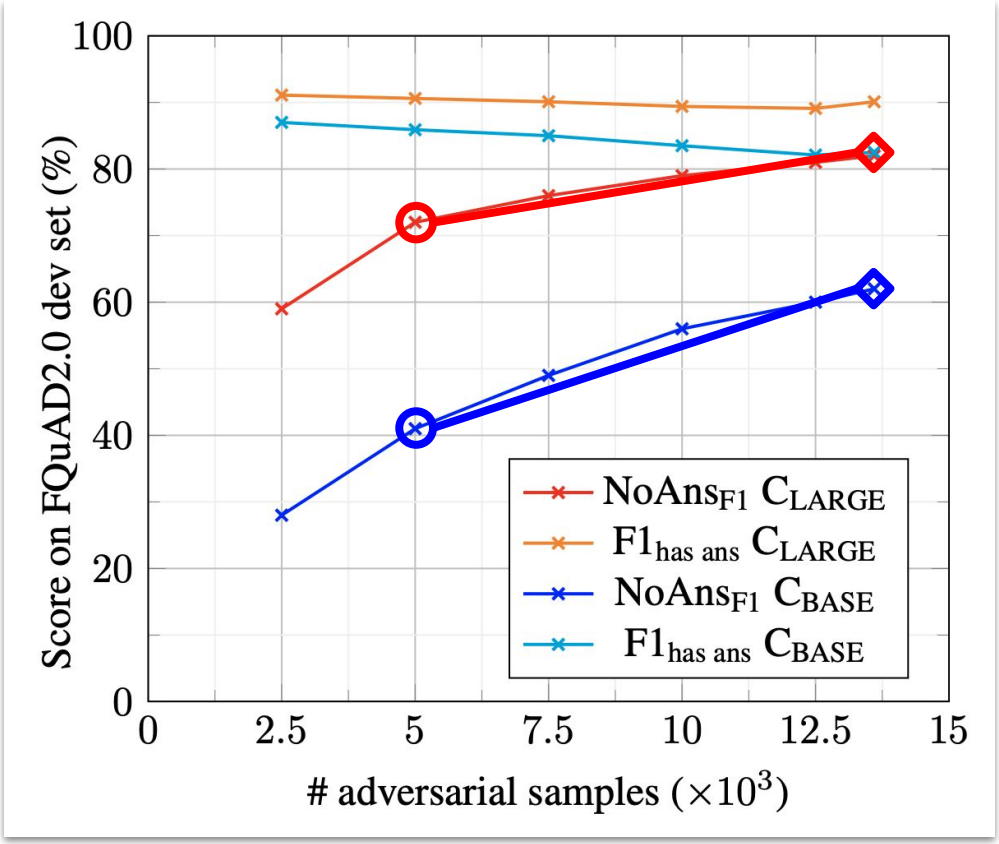


Figure 2: Evolution of NoAns$_{F1}$ and F1$_{has\ ans}$ for CamemBERT models depending on the number of unanswerable questions in the training dataset.

How many adversarial questions are needed for a model to learn to determine when a question is unanswerable?

➤ 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 are used and unanswerable questions are progressively added with increments of 2500 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.

French Monolingual Experiments

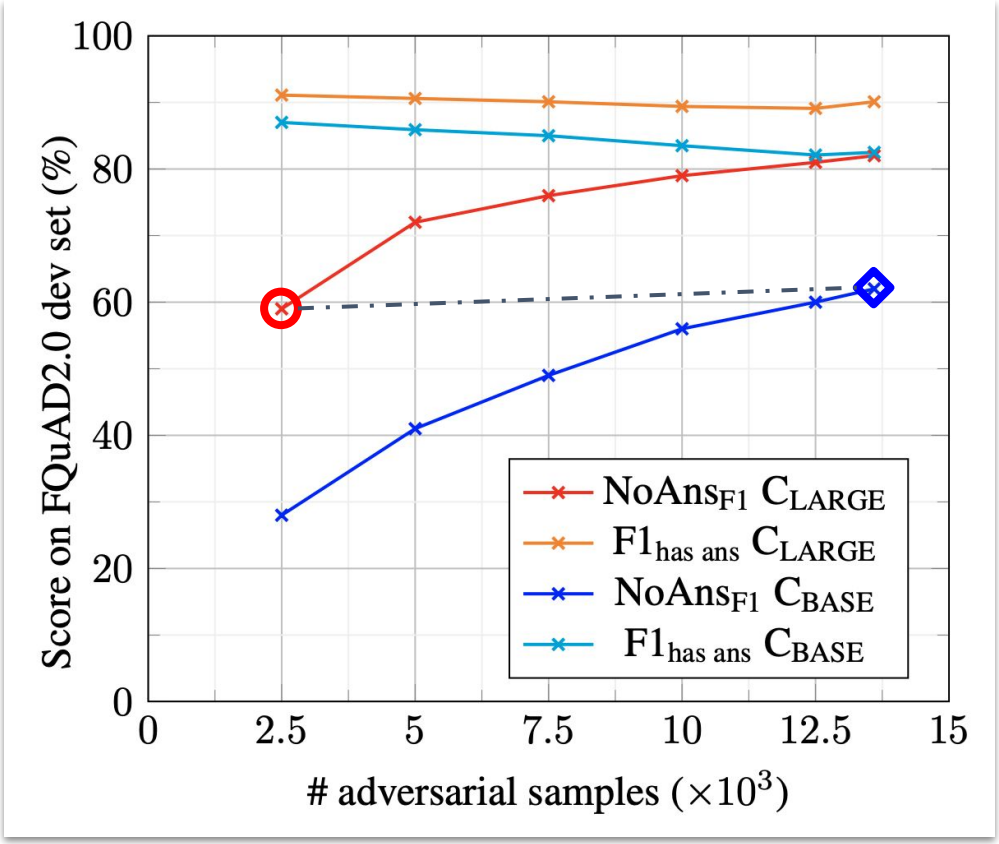


Figure 2: Evolution of $NoAns_{F1}$ and $F1_{has\ ans}$ for CamemBERT models depending on the number of unanswerable questions in the training dataset.

How many adversarial questions are needed for a model to learn to determine when a question is unanswerable?

➤ 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 are used and unanswerable questions are progressively added with increments of 2500 questions.

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 (5x more data).

French Monolingual Experiments

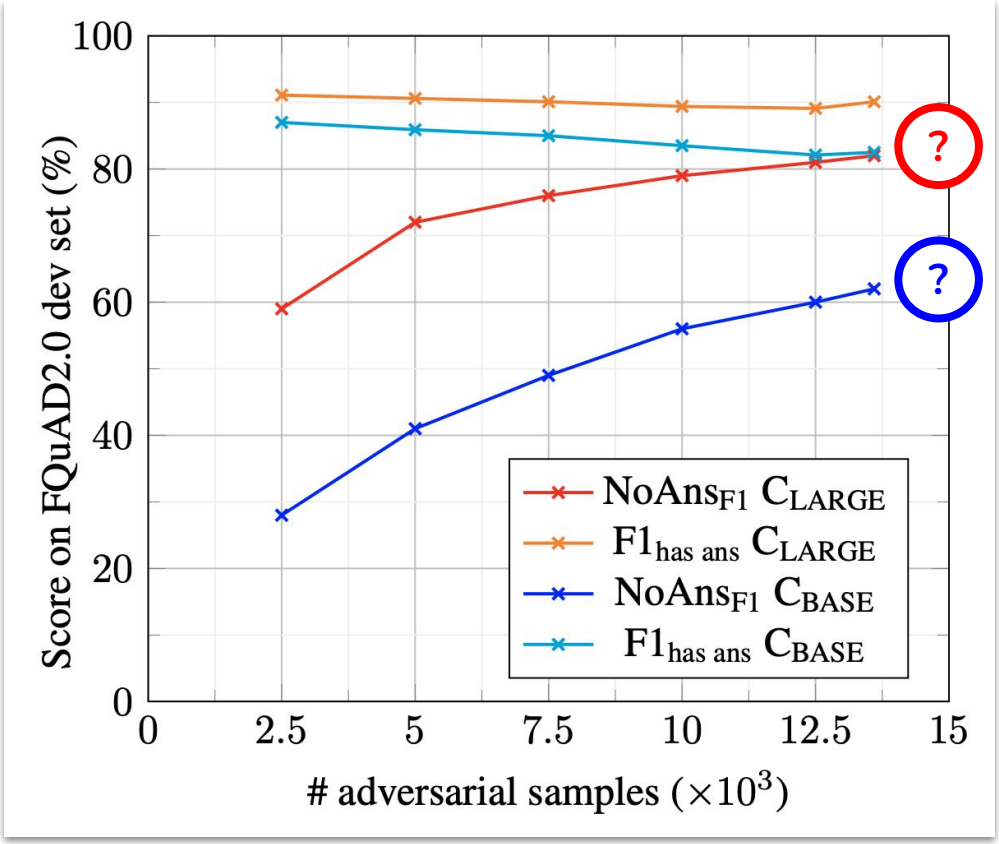


Figure 2: Evolution of $NoAns_{F1}$ and $F1_{has\ ans}$ for CamemBERT models depending on the number of unanswerable questions in the training dataset.

How many adversarial questions are needed for a model to learn to determine when a question is unanswerable?

➤ 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 are used and unanswerable questions are progressively added with increments of 2500 questions.

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

Multilingual Experiments

Could a multilingual Language Model fine-tuned solely on English Question Answering datasets compete against "FQuAD2.0", 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 on this Question Answering task?

Model	Training Dataset	Test Dataset	EM	F1	$F1_{has\ ans}$	$NoAns_{F1}$
$XLM\text{-}R_{BASE}$	SQuAD2.0	FQuAD2.0	56.0	62.4	75.9	56.2
	SQuAD2.0 + FQuAD2.0	FQuAD2.0	**64.4**	**69.6**	78.4	**66.4**
$CamemBERT_{BASE}$	FQuAD2.0	FQuAD2.0	63.3	68.7	**82.5**	62.1
$CamemBERT_{BASE}$	FQuAD2.0*	FQuAD2.0*	60.5	66.1	83.5	56.4
$RoBERTa_{BASE}$	SQuAD2.0*	SQuAD2.0*	69.7	73.3	85.3	73.4
$XLM\text{-}R_{LARGE}$	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
$CamemBERT_{LARGE}$	FQuAD2.0	FQuAD2.0	**78.0**	**83.0**	**90.1**	**82.3**

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

Multilingual Experiments

Results in zero-shot setting are promising: XLM-R_{LARGE} reaches in zero-shot setting better performances on the FQuAD2.0 dataset than CamemBERT$_{BASE}$ trained on FQuAD2.0.

Nevertheless, this observation must be put into perspective by reminding that the SQuAD2.0 training set includes 43.5k adversarial questions, hence 3.2 times more than FQuAD2.0.

Model	Training Dataset	Test Dataset	EM	F1	$F1_{has\ ans}$	$NoAns_{F1}$
XLM-R_{BASE}	SQuAD2.0	FQuAD2.0	56.0	62.4	75.9	56.2
	SQuAD2.0 + FQuAD2.0	FQuAD2.0	**64.4**	**69.6**	78.4	**66.4**
CamemBERT$_{BASE}$	FQuAD2.0	FQuAD2.0	63.3	68.7	**82.5**	62.1
CamemBERT$_{BASE}$	FQuAD2.0*	FQuAD2.0*	60.5	66.1	83.5	56.4
RoBERTa$_{BASE}$	SQuAD2.0*	SQuAD2.0*	69.7	73.3	85.3	73.4
XLM-R_{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$_{LARGE}$	FQuAD2.0	FQuAD2.0	**78.0**	**83.0**	**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.

Multilingual Experiments

For both model sizes, BASE and LARGE, the CamemBERT model reaches better performances than the XLM-R model in the zero-shot setting with a substantial margin.

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.

Model	Training Dataset	Test Dataset	EM	F1	$F1_{has\ ans}$	$NoAns_{F1}$
$XLM\text{-}R_{BASE}$	SQuAD2.0	FQuAD2.0	56.0	62.4	75.9	56.2
	SQuAD2.0 + FQuAD2.0	FQuAD2.0	**64.4**	**69.6**	78.4	**66.4**
$CamemBERT_{BASE}$	FQuAD2.0	FQuAD2.0	63.3	68.7	**82.5**	62.1
$CamemBERT_{BASE}$	FQuAD2.0*	FQuAD2.0*	60.5	66.1	83.5	56.4
$RoBERTa_{BASE}$	SQuAD2.0*	SQuAD2.0*	69.7	73.3	85.3	73.4
$XLM\text{-}R_{LARGE}$	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
$CamemBERT_{LARGE}$	FQuAD2.0	FQuAD2.0	**78.0**	**83.0**	**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.

Multilingual Experiments

By combining FQuAD2.0 and SQuAD2.0 training sets, XLM-R_{BASE} performs slightly better than CamemBERT$_{BASE}$ trained with FQuAD2.0, while for large models, CamemBERT is slightly better.

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.

Model	Training Dataset	Test Dataset	EM	F1	$F1_{has\ ans}$	$NoAns_{F1}$
XLM-R_{BASE}	SQuAD2.0	FQuAD2.0	56.0	62.4	75.9	56.2
	SQuAD2.0 + FQuAD2.0	FQuAD2.0	**64.4**	**69.6**	78.4	**66.4**
CamemBERT$_{BASE}$	FQuAD2.0	FQuAD2.0	63.3	68.7	**82.5**	62.1
CamemBERT$_{BASE}$	FQuAD2.0*	FQuAD2.0*	60.5	66.1	83.5	56.4
RoBERTa$_{BASE}$	SQuAD2.0*	SQuAD2.0*	69.7	73.3	85.3	73.4
XLM-R_{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$_{LARGE}$	FQuAD2.0	FQuAD2.0	**78.0**	**83.0**	**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.

Multilingual Experiments

By fine-tuning monolingual models on similar Question Answering datasets (50,000 answerable and 10,000 unanswerable questions), we find that all metrics are significantly better for the English set-up.

FQuAD2.0 is much harder than SQuAD2.0 in terms of both the difficulty and the ambiguity of questions. In particular, results for $NoAns_{F1}$ show that it is much harder for the French model to detect whether a question is answerable.

Model	Training Dataset	Test Dataset	EM	F1	$F1_{has\ ans}$	$NoAns_{F1}$
$XLM\text{-}R_{BASE}$	SQuAD2.0	FQuAD2.0	56.0	62.4	75.9	56.2
	SQuAD2.0 + FQuAD2.0	FQuAD2.0	**64.4**	**69.6**	78.4	**66.4**
$CamemBERT_{BASE}$	FQuAD2.0	FQuAD2.0	63.3	68.7	**82.5**	62.1
$CamemBERT_{BASE}$	FQuAD2.0*	FQuAD2.0*	60.5	66.1	83.5	56.4
$RoBERTa_{BASE}$	SQuAD2.0*	SQuAD2.0*	69.7	73.3	85.3	73.4
$XLM\text{-}R_{LARGE}$	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
$CamemBERT_{LARGE}$	FQuAD2.0	FQuAD2.0	**78.0**	**83.0**	**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.

Conclusion

Summary

- 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 non-English adversarial QA dataset**.
- 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.
- We showed the **superiority of a monolingual approach** on the target language using a dataset such as FQuAD2.0.
- We exhibited that **FQuAD2.0 is a harder dataset of adversarial QA than its English counterpart SQuAD2.0**.

Future work

- 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?**
- 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.
  - We could use smaller models than CamemBERT$_{LARGE}$ or CamemBERT$_{BASE}$, such as LePetit (Micheli et al., 2020)
  - 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).

Bibliographical References (1/3)

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

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

Brown, T. B.et al. (2020). Language models are few-shot learners.

Carrino, C. P., Costa-jussa`, M. R., and Fonollosa, J. A. R. (2019). 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.

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., Guzma ́n, 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.

d'Hoffschmidt, M., Belblidia, W., Heinrich, Q., Brendle ́, 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+).

Bibliographical References (2/3)

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.

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.

Martin, L., Muller, B., Ortiz Sua ́rez, P. J., Dupont, Y., Romary, L., Villemonte de la Clergerie, E., Seddah, D., and Sagot, B. (2019). CamemBERT: a Tasty French Language Model.

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/

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

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.

Bibliographical References (3/3)

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., Rockta ̈schel, 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.

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 Polo- sukhin, I. (2017). Attention is all you need. CoRR.

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.

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.