On the Usability of Transformers-based models for a French Question-Answering task

Oralie Cattan[1,2] **Christophe Servan**[1] **Sophie Rosset**[2]

[1]QWANT
61 rue de Villiers,
92200 Neuilly-sur-Seine, France
inital.lastname@qwant.com

[2]Université Paris-Saclay,
CNRS, LISN,
91405, Orsay, France
lastname@lisn.fr

Abstract

For many tasks, state-of-the-art results have been achieved with Transformer-based architectures, resulting in a paradigmatic shift in practices from the use of task-specific architectures to the fine-tuning of pre-trained language models. The ongoing trend consists in training models with an ever-increasing amount of data and parameters, which requires considerable resources. It leads to a strong search to improve resource efficiency based on algorithmic and hardware improvements evaluated only for English. This raises questions about their usability when applied to small-scale learning problems, for which a limited amount of training data is available, especially for under-resourced languages tasks. The lack of appropriately sized corpora is a hindrance to applying data-driven and transfer learning-based approaches with strong instability cases. In this paper, we establish a state-of-the-art of the efforts dedicated to the usability of Transformer-based models and propose to evaluate these improvements on the question-answering performances of French language which have few resources. We address the instability relating to data scarcity by investigating various training strategies with data augmentation, hyperparameters optimization and cross-lingual transfer. We also introduce a new compact model for French FrALBERT which proves to be competitive in low-resource settings.

1 Introduction

Recent advances in the field of Natural Language Processing (NLP) have been made with the development of transfer learning and the availability of pre-trained language models based on Transformer architectures (Vaswani et al., 2017), such as BERT (Devlin et al., 2019). As they provide contextualized semantic representation they contribute both to advance the state-of-the-art on several NLP tasks and also to evolve training practices through the use of fine-tuning.

The trend of recent years consists in training large pre-trained language models on ever larger corpora, with an ever-increasing amount of parameters, which requires considerable computational resources that only a few companies and institutions can afford. For example, the base model of BERT with 110 million parameters was pre-trained on 16 gigabytes (GB) of text, while the GPT-3 model (Brown et al., 2020) was pre-trained on 45 terabytes (TB) of text and has 175 billion parameters.

In fact, deploying ever-larger models raises questions and concerns about the increasing magnitude of the temporal, financial, and environmental cost of training and usability (Strubell et al., 2019; Moosavi et al., 2020). Typically, due to their resource requirements, these models are trained and deployed for industrial operations on remote servers. This leads to a high use of over-the-air communications, which are particularly resource-intensive (Gündüz et al., 2019). In particular, some NLP applications (speech recognition, speech to text, etc.) have some known problems related to network latency, transmission path difficulties, or privacy concerns. To reduce the impact of these communications, there is a solution that is to allow these models to run directly on peripheral or mobile devices, that is, in environments with limited resources that require lightweight, responsive models and energy efficiency. Reducing the size of the models is therefore one of the increasingly favoured avenues, especially for the reduction of memory resources and computation time involved in training and use.

To meet these constraints, compact models represent one of the most promising solutions. As far as we know, they have only been evaluated on the comprehension tasks covered by GLUE (Wang

Proceedings of Recent Advances in Natural Language Processing, pages 244–255
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_029

et al., 2018) and the question-answering task with the SQuAD corpus (Rajpurkar et al., 2016) with abundant data, in English. The improvements resulting from the algorithmic optimizations of the models, although significant, raise questions about their effectiveness on lower-scale learning problems on poorly endowed languages. The works of Zhang et al. (2021) and Mosbach et al. (2021) have furthermore shown degraded performance in these conditions. These two reflections are at the origin of a double question which our contribution attempts to answer. On the one hand, what is the behavior of a Transformer-based model in the context of a question-answering task in French, a task that is poorly endowed in this language? On the other hand, what are the impacts of algorithmic improvements of these same models in this context?

To answer these questions, we first establish in section 2 a state-of-the-art that is meant to be broad enough to have a shallow overview depicting the ins and outs and issues around the usability of Transformer-based models whose breadcrumb trail is the issue of resources. Then, we present in the section 3 the recent progress of the question-answering task, through the use of these latest models. In sections 4 and 5 we introduce our model and present our experiments on the usability of Transformers models in a question-answering task for French, on *FQuAD* (d'Hoffschmidt et al., 2020) and *PIAF* (Keraron et al., 2020) corpora. We propose to address the instability relating to data scarcity by investigating various training strategies with data augmentation, hyperparameters optimization and cross-lingual transfer. Finally, we present a new compact model for French based on ALBERT (Lan et al., 2020)[1], and compare it to existing monolingual and multilingual models, large and compact, under constrained conditions (notably on learning data).

2 Usability of Transformers

In this section we present the ins and out of the Transformer models to understand how the approaches meet the need for better usability.

2.1 Architecture and pre-trained models

The Transformer architecture (Vaswani et al., 2017) is based on a stack of encoder-decoder blocks, composed at a high level of forward propagation networks and multi-headed self-attention operations. The self-attention layer is the core element of its architecture that enables its efficiency in modeling the semantic context interdependencies between the units or sub-units of the input sequence.

Transformer-based language models such as BERT (Devlin et al., 2019) are pre-trained on large-scale data collections sourced from Wikipedia or Common Crawl (CC) with one or multiple training objectives (masked language modeling, next sentence or sentence order prediction). This pre-training can be followed by supervised fine-tuning according to the tasks, whether generatives (machine translation, abstractive summarization) or discriminatives (classification, question-answering). The ensuing fine-tuning phase allows for better initialization of the models parameters while requiring less task-specific data so as to make the training of subsequent tasks faster.

Recently, Zhang et al. (2021) and Mosbach et al. (2021) have nevertheless shown that the commonly adopted practices (the number of iterations, the choice of model layers) when fine-tuning Transformers-based langage models are inappropriate under resource constrained conditions and adversely affect the stability of models performances as overfitting, label noise memorization or catastrophic forgetting. Added to this, because the pre-training process is particularly constraining, various works have been oriented towards the research and training of efficient models, both in terms of available capacities and resources and in terms of environmental footprint.

2.2 A search for efficiency

Reducing the cost of training Transformers-based models has become an active research area. To this end, methods based on compression techniques or on architecture improvements have been introduced in order to build compact models with comparable performances to large models.

Many works address the issue of model compression with quantization, pruning, knowledge distillation or a combination of these approaches. The idea of quantization (Shen et al., 2020) is to take advantage of the use of lower precision bit-width floats to reduce memory usage and increase computational density. Following the same objective, pruning (Michel et al., 2019) consists in removing parts of a model (weight bindings, attentional heads) with minimal precision losses. Finally, knowledge distillation (Sanh et al., 2019) enables the generation

[1] Available at HuggingFace's model hub page.

of models that mimic the performance of a large model (or set of models) while having fewer parameters.

Another axis of development concerns the use of neural architecture search (Elsken et al., 2019) which allows to optimize a model by progressively modifying the design of the network through trial and error, eliminating insignificant operations. To avoid the unnecessary large number of parameters, adapters (Houlsby et al., 2019) were introduced to allow fine-tuning of the set of parameters specific to the task of interest rather than the entire model.

Other architectural improvements highlighted with the introduction of the ALBERT model (Lan et al., 2020) such as the factorization of the attention matrix or parameter sharing. Indeed, the most time-consuming and memory-intensive operations concerns the forward propagation and attention computation operations. The self-attention layer of BERT pretrained models grows quadratically in respect to the input sequence length. One common approach to this issue consists of approximating the dot-product attention for example by using hashing techniques (Kitaev et al., 2020) to accelerate the training and inference phases when long sequence lengths are used. However these solutions have demonstrated they suffer from important computational overheads for tasks with smaller lengths, such as question-answering.

3 The Question-Answering task

Question-Answering (QA) based on machine reading comprehension corresponds to the task of extracting an answer given a question and a context document such as from a news or Wikipedia article.

3.1 General QA Architecture

Until recently, most of the proposed approaches have relied on an architectural complexification of LSTM-based neural networks and attention mechanism. At a high level, their architectures are all composed of three layers with:

(a) **an encoding layer** that projects the inputs, as each word within the context, question and answer triples in a latent semantic space;

(b) **an interaction layer** that models the semantic interdependencies between the embedded inputs through the use of attention mechanisms.

(c) **an output layer** that extracts the answer to the input question within the related context.

The interaction layer is the core element of the architecture for which several kinds attention mechanisms has been developed to improve the QA matching process such as bi-attention (Seo et al., 2017), co-attention (Xiong et al., 2017, 2018), multi-level inter-attention (Huang et al., 2018) or re-attention (Hu et al., 2018), to name just a few.

Recent advances through the availability of Transformer-based pre-trained models and the development of transfer learning methods have enabled to remove the recurrence of previous architectures in order to achieve parallelization efficiencies. This simplified the QA architecture and its training process, replacing the encoding and the interaction layers with attention-based Transformer layers.

Another advantage is that it provides pre-computed contextual word representations. QA models based on LSTMs are built on top of static word embeddings models such as GloVe (Pennington et al., 2014). Even these models have up to 40 times fewer parameters than a BERT-based model, they rely on LSTM-based encoders to produce contextual embeddings which considerably lengthens the time required for training and makes the dependence on supervised data more important.

The standard approach introduced by Devlin et al. (2019) we rely on this study, consists in introducing and updating parameter vectors corresponding to the start and end positions of the answer span. Specifically, the start and end position probability distributions are computed by softmax over the dot products between the representation of the tokens and the start and end vectors. In sum, all of the Transformer parameters as well as the two introduced parameter vectors are optimized together.

Despite the fact that there is a number of large-scale QA datasets in English, with tens of thousands of annotated training examples, porting a system to a new language with fewer annotated resources (low-resource languages) requires approaches that go far beyond the simple act of retraining the models.

3.2 Low-Resourced QA

In recent years, low-resource NLP has drawn an increasing amount of attention with solutions ranging from developing new data collection methodologies either via crowdsourcing or through the use of machine translation (MT), to cross-lingual and transfer learning approaches for which information is shared across languages or tasks.

3.2.1 MT-based data collection

Neural MT as made considerable progress in recent years such as translating large-scale datasets from a high-resourced to under-resourced languages or converserly has become an intuitive way of generating annotated datasets in a cost-effective and rapid manner.

Automatically translating the context, question and answer triples from a high-resource language, such as English (called source domain) to low-resource languages (called target domains) have enabled the evaluation of models for languages with no training data available but also the creation of large-scale MT-based QA corpora for the Italian (Croce et al., 2018), Spanish (Carrino et al., 2020), Arabic (Mozannar et al., 2019) and Korean (Youngmin Kim, 2020) languages.

Another approach consists of translating the QA triples of the target domain into the source domain, so the model trained on the source language can be directly applied on the translated target language testing data. As an exemple, Asai et al. (2018)'s method consisted of combining the alignment attention scores from a MT model with an English QA model to guide the answer extraction process.

The performance of MT-data approaches depends strongly on the quality of the MT models. Thus, due to the lack of reliable models for some language pairs, approaches that foster the transfer of knowledge from other languages or tasks while requiring less data have been developed.

3.2.2 Pre-training and Transfer approaches

The exploitation of pre-trained models followed by task-specific fine-tuning haved pushed the state-of-the-art forwards, while requiring much less computational and data resources. The idea behind pre-training is to reuse the weights parameters trained on a set of source tasks and continue to fine-tune them on under-resourced target tasks to achieve knowledge transfer. Dai and Le (2015) were the first to propose to pre-train RNNs using auto-encoders and language models as part of their QA encoding layer. Min et al. (2017) and Wiese et al. (2017) pre-trained QA models before applying the fine-tuning process between the source and the target domains. Other efforts focused on pre-training Transformer-based models multilingually such as the multilingual version of BERT (called mBERT) (Devlin et al., 2019) or XLM-R (Conneau et al., 2020) to learn cross-lingual representations which are transferable across languages.

3.2.3 Usability concerns

Studies on the usability of Transformer-based models (Section 2) from standard resource efficiency concerns towards a broader set of problems related to their generalizability.

Recently, Pires et al. (2019) and Conneau et al. (2020) have shown that multilingual models underperformed, when applied on poorly endowed languages. Additionally, as mentioned in subsection 2.1, recent works (Zhang et al., 2021; Mosbach et al., 2021) have highlighted the limitation of Transformer-based transfer learning with strong instabilities arising from the small-scale learning.

French is a poorly endowed language since we do not have enough annotated data to train a deep learning model on QA tasks. Moreover, unlike the only two large monolingual French models: CamemBERT (Martin et al., 2020) and FlauBERT (Le et al., 2020), the English BERT model has become a branching point from which a growing number of large and compact English pre-trained models have emerged. These French monolingual models, although they provide good performances, do not reflect the rapid evolution of the field.

Consequently, in this paper we propose a new compact model FrALBERT for French we present in the following section alongside with the other available pre-trained language models on which we base our experiments.

4 FrALBERT and Transformer models considered

As mentioned in the previous section there is no compact model for French. We therefore decided to pre-train a new version of ALBERT from scratch we called FrALBERT, thus overcoming some of the discussed limitations.

ALBERT is based on parameter sharing/reduction techniques that allows to reduce the computational complexity and speed up training and inference phases. Compared to previous compact models such as DistilBERT (Sanh et al., 2019), Q-BERT (Shen et al., 2020) or TernaryBERT (Zhang et al., 2020), ALBERT is to the date the smallest pre-trained models with 12 million parameters and <50 megabyte (MB) model size.

FrALBERT is pre-trained on the French version of the Wikipedia encyclopedia of 04/05/2021, i.e. 4 GB of text and 17 million (M) sentences. Beyond concerns about the rights to use data from Common

model	pre-training data	vocab. size	# param.	model size
CamemBERT$_{\text{base}}$	French OSCAR (138 GB of text)	32005	110 M	445 MB
CamemBERT$_{\text{large}}$	French CCNet (135 GB of text)	32005	335 M	1.35 GB
CamemBERT$_{\text{base}}$	French Wikipedia (4 GB of text)	32005	110 M	445 MB
FrALBERT$_{\text{base}}$	French Wikipedia (4 GB of text)	32005	12 M	50 MB
XLM-R$_{\text{base}}$	CC-100 (2.5 TB of text)	250002	278 M	1.12 GB
XLM-R$_{\text{large}}$	CC-100 (2.5 TB of text)	250002	559 M	1.24 GB
mBERT$_{\text{base}}$	Wiki-100	119547	177 M	714 MB
small-mBERT$_{\text{base}}$	Wiki-100	33407	111 M	447 MB
distil-mBERT$_{\text{base}}$	Wiki-100	119547	134 M	542 MB

Table 1: Characteristics of the pre-trained models used in the proposed small-scale QA framework.

Crawl projects such as OSCAR (Ortiz Suárez et al., 2020) or CCNet (Wenzek et al., 2020) corpora, and because we focus on factual QA, we decide to use only Wikipedia as our primary source of knowledge. We used the same learning configuration as the original model with a batch size of 128 and a initial learning rate set to 3.125 × 10^{-4}.

Our experiments are also based on the large monolingual French model CamemBERT (Martin et al., 2020) as well as on the two large multilingual models: XLM-R (Conneau et al., 2020) and mBERT (Devlin et al., 2019), both pre-trained from massive corpora dataset in more than 100 languages such as the Common Crawl (CC-100) or Wikipedia (Wiki-100). We also exploit two compact multilingual models with a distilled version of mBERT: distil-mBERT (Sanh et al., 2019) and small-mBERT (Abdaoui et al., 2020), a mBERT model whose the original vocabulary has been reduced to two languages (English and French). Table 1 gives a comparison of the models.

5 Experiments

We propose an assessment of the comparative advantage gains in performance when using different training strategies (data augmentation, hyperparameter search and cross-lingual transfer) over monolingual and multilingual pre-trained models, larges and compacts for a QA task in French under resource constraints.

5.1 QA Datasets

We conduct experiments on four QA datasets whose descriptives are presented in Table 2, with:

- *SQuAD* (v1.1) (Rajpurkar et al., 2016) (we called *SQuAD-en*) the reference corpus to evaluate QA models' performances in English, consisting of 100K+ QA pairs sourced from 442 English Wikipedia articles;

- *FQuAD* (v1.0) (d'Hoffschmidt et al., 2020), a recently released French QA dataset consisting of 25K+ crowdsourced QA pairs based on 135 articles on French Wikipedia;

- *PIAF* (v1.0) (Keraron et al., 2020), a small-scale dataset in French with only 3K+ pairs of QA pairs in 191 Wikipedia articles; and

- *SQuAD-fr*$_{\text{train}}$, our French translated version of *SQuAD-en*. We used the Transformer architecture as described in Vaswani et al. (2017) from the Open NMT framework (Klein et al., 2017) (Open-Source Neural Machine Translation) implementation of the network to train our neural MT system. When translating QA corpora, the problem we face is that the translated answer may not be present in the translated context. Thus, simple techniques such as segment matching are inadequate to retrieve the answer. We have developed an answer extraction process that is based on

dataset	*SQuAD-en*$_{\text{train}}$	*FQuAD*$_{\text{train}}$	*FQuAD*$_{\text{dev}}$	*PIAF*$_{\text{dev}}$
# titles	442	117	18	191
# paragraphs	18,891	4,920	768	761
# sentences / # tokens / #characters	5.0 / 119.7 / 635.6	4.8 / 125.9 / 653.8	5.4 / 147.3 / 765.7	4.2 / 110.5 / 579.1
# questions	87,342	20,088	3,184	3,812
# tokens / # characters	10.1 / 50.2	9.2 / 47.9	8.5 / 45.6	9.1 / 48.2
# answers	87,599	20,731	3,188	3,835
# tokens / # characters	2.9 / 16.8	3.6 - 19.5	4.3 - 23.0	4.5 - 24.1
human performance (F1 / EM)	90.5 / 80.3	92.6 / 79.5		-

Table 2: Descriptives of *SQuAD-en*, *FQuAD* and *PIAF* datasets.

ChrF (Popović, 2015) a character n-gram precision and recall enhanced with word n-grams. Since answers are largely made up of entities, ChrF score integration is only performed when the answer span is not present in the related context. In order to evaluate the quality of the translation, we manually corrected the translation errors in the output of a subset of the corpus composed of 890 QA pairs and 107 contexts. We obtain a BLEU score (Papineni et al., 2002) of 68.89 and 72.38 for questions and contexts respectively. *SQuAD-fr*$_{\text{train}}$ serves as a means of data augmentation on *FQuAD* and *PIAF* benchmarks, with 90K+ translated QA training pairs[2].

We also explore mixed datasets training strategy with *SQuAD-en*$_{\text{train}}$ + *FQuAD*$_{\text{train}}$ for training models on a concatenation of the training data covering French-English language pairs to test the cross-lingual transfer ability of multilingual models.

5.2 Evaluation and validation

The performance of QA models are evaluated using the Exact Match (EM) and F1 scores. The EM score is the percentage of system outputs that match exactly with the ground truth answers. The F1 score is a combined measure of precision and recall that is less strict than EM. The evaluation process[3] involves post-processing identical to that presented by d'Hoffschmidt et al. (2020) and inspired by that proposed for English by Rajpurkar et al. (2016), which consists of the removal of punctuation marks and determiners[4] as well as a down-casing of the answers (ground truths and predictions).

To address our considerations related to resource constraints we perform a hyperparameter optimization, that has proven to lead to better solutions in less time. It is based on a population-based learning (Jaderberg et al., 2017) in which a population of models and their hyperparameters are jointly optimized. To this end, we build a validation set by randomly extracting 10% of the training data.

5.3 Results

Table 3 presents the results on the French QA task evaluated on *FQuAD*$_{\text{dev}}$ and *PIAF*$_{\text{dev}}$. This table shows the scores obtained with Transformer-based models on the baseline training (*FQuAD*$_{\text{train}}$), using hyperparameter optimization approach (*FQuAD*$_{\text{train}}$ w/ optim) and with data augmentation approach (*FQuAD*$_{\text{train}}$ + *SQuAD-fr*$_{\text{train}}$). Cross-lingual performances are presented in table 4, on the same French QA tasks *FQuAD* and *PIAF* as baseline the English corpus *SQuAD-en*$_{\text{train}}$, on which we applied hyperparameter optimization (*SQuAD-en*$_{\text{train}}$ w/ optim) and performed data augmentation by adding the French corpus *FQuAD* to the English QA training corpus (*SQuAD-en*$_{\text{train}}$ + *FQuAD*$_{\text{train}}$).

5.3.1 Baseline results

The results obtained from monolingual models with CamemBERT$_{\text{large}}$ which has more layers, hidden units and attention heads and Camembert$_{\text{base}}$, both pre-trained with a larger and more diverse amount of data achieve results are better than Camembert$_{\text{base}}$ pre-trained on only 4 GB Wikipedia. The highest F1 score is 81.2 on *FQuAD*$_{\text{dev}}$ and 68.1 on *PIAF*$_{\text{dev}}$.

The F1 performances of the FrALBERT$_{\text{base}}$ model are close to those of the CamemBERT$_{\text{base}}$ model, both pre-trained on the French content of Wikipedia (4GB). Their results turn out to be competitive and of the same order of magnitude as those reported by Lan et al. (2020) on *SQuAD-en* with 1 point difference on F1 scores when evaluating a BERT$_{\text{base}}$ model (90.4 F1) and a compact ALBERT$_{\text{base}}$ model (89.3 F1) pre-trained on the same texts (BookCorpus and Wikipedia). Interestingly, these EM scores are higher than those of the CamemBERT$_{\text{base}}$ achieving the EM score of 55.1, an increase of 5 points on the FQuAD$_{\text{dev}}$.

5.3.2 hyperparameter results

Automatically tuning the hyperparameter tends to make QA models more accurate with gains in terms of EM scores that are very expressive. Highest F1 / EM scores are 90.2 / 75.5 on *FQuAD*$_{\text{dev}}$ and 71.0 / 44.8 on *PIAF*$_{\text{dev}}$. Improvement are variable accoring the model considered, especially the French BERT one wich have the highest improvement using this approach (from 6 to 9 F1 points and from 11 to 20 EM points) which is quite imrpessive. FrALBERT stay behind of 5 F1 points of the CamemBERT$_{\text{base}}$ trained with the same data (wiki

[2] We share our *SQuAD-fr* corpus on request and on dataset sharing platforms to support further research in this area.

[3] The experiments reported in (d'Hoffschmidt et al., 2020) concern version 1.1 of the *FQuAD* corpus. Ours are based on the only version available to date (v1.0). Moreover, the test sets are not made public, so we use the development set instead.

[4] Determiners are *le, la, les, l', du, des, au, aux, un, une*.

testing data	***FQuAD*$_{\text{dev}}$**			***PIAF*$_{\text{dev}}$**		
model \ training strategy	***FQuAD*$_{\text{train}}$**	***FQuAD*$_{\text{train}}$ w/ optim.**	***FQuAD*$_{\text{train}}$ + *SQuAD-fr*$_{\text{train}}$**	***FQuAD*$_{\text{train}}$**	***FQuAD*$_{\text{train}}$ w/ optim.**	***FQuAD*$_{\text{train}}$ + *SQuAD-fr*$_{\text{train}}$**
CamemBERT$_{\text{base}}$	77.6 / 52.5	85.5 / 70.3	**86.7 / 71.7**	62.0 / 37.5	63.8 / 38.9	**64.3 / 39.2**
CamemBERT$_{\text{large}}$	81.2 / 55.9	**90.2 / 75.5**	89.9 / 75.2	68.1 / 42.2	**71.0 / 44.8**	68.9 / 42.5
CamemBERT$_{\text{base (wiki 4 GB)}}$	74.2 / 49.5	80.7 / 61.8	**85.1 / 69.5**	61.7 / 37.3	62.9 / 37.9	**65.9 / 41.0**
FrALBERT$_{\text{base (wiki 4 GB)}}$	72.6 / 55.1	75.6 / 64.8	**84.3 / 70.5**	61.0 / 38.9	62.1 / 39.5	**66.9 / 43.7**
XLM-R$_{\text{base}}$	82.1 / 66.8	83.1 / 67.9	**84.2 / 68.8**	65.0 / 39.6	66.9 / 41.2	**68.6 / 42.7**
XLM-R$_{\text{large}}$	86.8 / 71.5	**89.5 / 75.8**	87.3 / 72.5	70.4 / 43.8	**73.2 / 45.8**	72.6 / 45.2
mBERT$_{\text{base}}$	78.6 / 61.8	82.5 / 65.7	**84.1 / 68.6**	62.5 / 37.8	64.1 / 38.0	**64.8 / 40.0**
small-mBERT$_{\text{base}}$	75.1 / 55.7	78.0 / 62.2	**81.6 / 64.6**	60.8 / 35.6	62.2 / 37.7	**63.7 / 39.8**
distil-mBERT$_{\text{base}}$	72.8 / 56.0	73.0 / 55.1	**78.1 / 61.5**	52.3 / 30.1	53.6 / 31.4	**58.3 / 34.9**

Table 3: Results obtained with French and multilingual Transformer models on the baseline training (*FQuAD*$_{\text{train}}$), using hyperparameter optimization (*FQuAD*$_{\text{train}}$ w/ optim) and with data augmentation (*FQuAD*$_{\text{train}}$ + *SQuAD-fr*$_{\text{train}}$) using F1-measure (F1) and Exact Match (EM), on two French QA tasks (*FQuAD*$_{\text{dev}}$ and *PIAF*$_{\text{dev}}$).

testing data	***FQuAD*$_{\text{dev}}$**			***PIAF*$_{\text{dev}}$**		
model \ training strategy	***SQuAD-en*$_{\text{train}}$**	***SQuAD-en*$_{\text{train}}$ w/ optim.**	***SQuAD-en*$_{\text{train}}$ + *FQuAD*$_{\text{train}}$**	***SQuAD-en*$_{\text{train}}$**	***SQuAD-en*$_{\text{train}}$ w/ optim.**	***SQuAD-en*$_{\text{train}}$ + *FQuAD*$_{\text{train}}$**
XLM-R$_{\text{base}}$	81.3 / 65.0	82.5 / 66.5	**83.6 / 67.5**	61.4 / 37.2	62.7 / 38.5	**64.9 / 39.9**
XLM-R$_{\text{large}}$	82.8 / 64.8	84.4 / 67.8	**87.1 / 72.0**	65.1 / 39.1	66.3 / 40.5	**69.0 / 43.2**
mBERT$_{\text{base}}$	76.0 / 59.3	79.5 / 62.3	**83.5 / 67.6**	61.6 / 37.2	62.1 / 36.9	**64.5 / 39.6**
small-mBERT$_{\text{base}}$	73.1 / 49.0	76.0 / 59.1	**81.4 / 62.1**	59.6 / 36.5	61.0 / 37.8	**63.0 / 38.9**
distil-mBERT$_{\text{base}}$	65.4 / 47.4	68.6 / 48.5	**75.9 / 56.3**	48.8 / 28.1	52.0 / 29.2	**56.5 / 33.1**

Table 4: Cross-language transfer results obtained with multilingual Transformer models only on the baseline (*SQuAD-en*$_{\text{train}}$), using hyperparameter optimization (*SQuAD-en*$_{\text{train}}$ w/ optim) and with data augmentation (*SQuAD-en*$_{\text{train}}$ + *FQuAD*$_{\text{train}}$) using F1-measure (F1) and Exact Match (EM), on two French QA tasks (*FQuAD*$_{\text{dev}}$ and *PIAF*$_{\text{dev}}$).

4 GB) but regarding the EM scores, FrALBERT is better of 3 points. Surprisingly, multilingual models are close to the French BERT models. The small-mBERT is better than FrALBERT around 2.5 F1 points, while the French one have a better EM score (+2.6 points), while distil-mBERT is lower in both F1 and EM scores.

5.3.3 Data augmentation results

Training strategies based on data augmentation got nearly the best results in both F1 and EM scores except for the CamemBERT$_{\text{large}}$. Apart from CamemBERT$_{\text{base (wiki 4 GB)}}$ and FrALBERT$_{\text{base}}$ models, results are comparable with an average difference of 1 point regardless of the metric. More generally, the performance gains are up to 11 and 20 of F1 and EM points, respectively, on *FQuAD*$_{\text{dev}}$ and up to 4 points on both metrics on *PIAF*$_{\text{dev}}$.

5.3.4 Cross-lingual transfer results

The cross-lingual transfer-based approaches using multilingual models outperform the monolingual approaches on *FQuAD* and *PIAF* corpora (table 4). Once again the large model XLM-R achieves better results than its base version. XLM-R pretrained with a dual LM objective lens scores better than the mBERT model every time. The highest F1 score is 86.8 on *FquAD*$_{\text{dev}}$ and 70.4 on *PIAF*$_{\text{dev}}$. There is a significant performance drop between the large multilingual models and their respective compact models. The compact multilingual models based on mBERT substantially underperform, obtaining lower F1 and EM scores than the large models regardless of the training strategy. In the zero-shot configurations where no French data is used for training (*SQuAD-en*$_{\text{train}}$ and *SQuAD-en*$_{\text{train}}$ w/ optim), the models confirm the outstanding crosslingual ability with performances exceeding the performances of the monolingual models with F1 scores with an F1 / EM scores slightly below those obtained on *FQuAD*$_{\text{train}}$.

5.3.5 General observations

In all configurations, the performance in terms of EM and F1 on PIAF remains significantly lower than that obtained on *FQuAD* since the *PIAF* corpus does not include multiple responses as pointed out by d'Hoffschmidt et al. (2020). Unsurprisingly, *PIAF*$_{\text{dev}}$ offer a more challenging evaluation set, where the answer extraction performance are lower. Indeed, the corpus is more diversified with questions on 191 different Wikipedia articles, whereas on *FQuAD*$_{\text{dev}}$ it only covers 18.

According results, we can confirm that data augmentation is the better way to improve results, even if data comes from another language. We observe from multiligual results that combining training data gives better results with similar performance whether data are translated or not.

5.4 Analysis

In this section, we conduct an analysis of our results to understand what remains as challenges for state-of-the-art models, with a focus on the usability concerns of Transformers under resource constraints.

The success of supervised methods depends heavily on the availability of large-scale training data. Pre-training large models on massive corpora using unsupervised language modeling and fine-tuning the model with pre-trained weights requires less task-specific data. Our experimental results are in line with this, since we obtain satisfactory results with transfer learning when large and high quality annotated data are not available. The highest F1 score is 81.2 on *FQuAD*$_{\text{dev}}$ and 68.1 on *PIAF*$_{\text{dev}}$. Nevertheless, the performance of the models can benefit from several training strategies.

Effect of MT-data augmentation The lack of human-annotated datasets for languages other than English can be overcome by enriching our training data with the translated version of *SQuAD-en*$_{\text{train}}$. Regardless of the pre-trained model used, their performance are competitive on *FQuAD*$_{\text{dev}}$ and *PIAF*$_{\text{dev}}$, close to human performance.

Effect of hyperparameter tuning A generally unstated assumption is that pre-trained linguistic models are under-optimized and that practices commonly adopted for the fine-tuning stage can be detrimental to performance (Zhang et al., 2021; Mosbach et al., 2021). This is quite apparent in all settings, with better gains through hyperparameter optimization stages. Fine-tuning CamemBERT$_{\text{large}}$ on the French dataset yields 90.2 / 75.5 F1 / EM on the *FQuAD* dev set. By means of comparison, CamemBERT$_{\text{large}}$ scores were 81.2 / 55.9 F1 / EM on the same set with no hyperparameter tuning.

Crosslingual QA Pre-training language models on the concatenation of multiple languages has proven to be a competitive approach for cross-lingual language modeling. If monolingual models often perform better than multilingual models, we observe that, for comparable model sizes this is not the case in our task where the performance gap is smaller. This gap is further reduced when the strategies are combined. Scores of fine tuned XLM-R$_{\text{base}}$ is 82.1 / 66.8 on *FQuAD*$_{\text{dev}}$ and 65.0 / 39.6 on *PIAF*$_{\text{dev}}$.

The zero-shot experiments show that multilingual models can reach strong performances on the task in French when the model has not encountered data of the French language. For example, the XLM-R$_{\text{base}}$ model fine-tuned solely on SQuAD-en$_{\text{train}}$ reaches a performance on FQuaD just a few points below the performance obtained when fine-tuning is performed on FQuAD$_{\text{train}}$.

Finally, our results suggest that data-driven augmentation, either by translating datasets from high resource languages or by concatenating the available corpora are a particularly appropriate strategy to exploit the potential of cross-lingual transferability of models and data for improving model performances.

Improvements over a small scale dataset With resource-limited training data we obtain an average F1 score of 74.2 on *FQuAD*$_{\text{dev}}$ and 61.7 on *PIAF*$_{\text{dev}}$ when we fine-tuned FrALBERT — higher performance than any of the compact multilingual models, but slightly below the performance of the large monolingual models. We believe that the lower performance of small multilingual models is not due to their lower number of parameters, but to the usability of these models which is dependent on the reduction technique used. This can be seen very clearly since the performance of the FrALBERT model is close to that of the large models.

Here again, these performances can be boosted via the use of translated data or hyperparameter search which allows us to bring the maximum performances obtained with FrALBERT$_{\text{base}}$ and CamemBERT$_{\text{base}}$ pre-trained on 4 GB Wikipedia closer in a consistent way. Their scores remain slightly below those obtained with models pre-trained on more data suggesting limitations related to the corpus domain of the language model.

Computational costs Compact models provide alternatives to high-energy consumption models by showing comparable performance while reducing their size and computational complexity. Decreasing the environmental impact of NLP model training, as a research topic, is very recent (Moosavi et al., 2020). We decided to monitor our experiments conducted on one NVIDIA V100 GPU with 16GB of memory using

model	# param.	model size	Time (s)	Energy (kWh)	CO$_2$ (g)
CamemBERT$_{\text{base}}$	110 M	445 MB	7,207	1.08	317.87
CamemBERT$_{\text{large}}$	335 M	1.35 GB	19,445	3.10	914.27
FrALBERT$_{\text{base}}$	**12 M**	**50 MB**	**3,816**	**0.57**	**167.80**
XLM-R$_{\text{base}}$	278 M	1.12 GB	7,676	1.14	337.70
XLM-R$_{\text{large}}$	559 M	1.24 GB	21,137	3.30	973.29
mBERT$_{\text{base}}$	177 M	714 MB	7,333	1.07	317.02
small-mBERT$_{\text{base}}$	111 M	447 MB	7,190	1.09	321.42
distil-mBERT$_{\text{base}}$	134 M	542 MB	6,466	1.06	314.17

Table 5: Comparison of models by computational costs on *FQuAD*$_{\text{train}}$

`experiment-impact-tracker` (Henderson et al., 2020). Table 5 shows the energy consumption in kilowatt-hour (kWh), the emission intensity in grams of carbon dioxide (g CO$_2$), and the duration in seconds (s) for a fine tuning session of 10 epochs with a batch size of 4 on *FQuAD*$_{\text{train}}$.

The footprint of the large versions of XLM-R$_{\text{large}}$ and CamemBERT$_{\text{large}}$ models is 3 times more than their base versions. Their training time is also significantly longer, over 5 hours. Finally, in terms of watt usage, carbon emissions and training time, FrALBERT is two times less the distilled version of BERT.

6 Conclusion and outlook

Recently, important progress has been made in neural language modeling using Transformer networks. Its popularity now well established lies in its effectiveness in modeling long-term dependencies. In this study, we have shown that a number of significant shortcomings of usability have recently been pointed out and that some solutions have been drawn up with compact models. We have also overviewed how the use of Transformer-based pretrained language models have sparked a paradigmatic shift in question-answering training practices from task-specific architectures to the use of transfer learning through fine-tuning.

Comparing performances on a French question-answering task using large and compact models provides insight into the usability of these models for under-resourced languages. As others, we argued that large and compact models cannot be used with limited data. Our experimental results suggest that training strategy such as hyperparameter tuning or data augmentation can help to alleviate the data-gathering burden, with performances close to those of a high-resourced language such as English.

Finally, we present a new compact model for French FrALBERT (12M parameters), which proves to be as competitive as the large monolingual model CamemBERT (110M parameters) pretrained on the same amount of text. In term of computational cost, we shown this compact model is twice less greedy than the BERT$_{\text{base}}$ models. We also release a high-quality translated version of the SQuAD corpus in French consisting of around 90K+ QA pairs.

In a future work, we aim to continue this study from a meta-learning perspective with a model-agnostic approach generalizable to low-resource languages. We also plan to extend our model to other languages and to evaluate it on other NLP tasks such as named entity recognition or natural language understanding.

Acknowledgments

This work has been partially funded by the French National Research Agency (ANR) through the project TextToKids (AAPG2019).

References

Amine Abdaoui, Camille Pradel, and Grégoire Sigel. 2020. Load what you need: Smaller versions of mutililingual BERT. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 119–123, Online. Association for Computational Linguistics.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Casimiro Pio Carrino, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Automatic Spanish translation of SQuAD dataset for multi-lingual question answering. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5515–5523, Marseille, France. European Language Resources Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2018. Neural learning for question answering in italian. In *AI*IA 2018 – Advances in Artificial Intelligence*, pages 389–402, Cham. Springer International Publishing.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.

Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. 2019. Neural architecture search: A survey. *Journal of Machine Learning Research*, 20(55):1–21.

Deniz Gündüz, Paul de Kerret, Nicholas D. Sidiropoulos, David Gesbert, Chandra R. Murthy, and Mihaela van der Schaar. 2019. Machine learning in the air. *IEEE Journal on Selected Areas in Communications*, 37(10):2184–2199.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. 2018. Reinforced mnemonic reader for machine reading comprehension. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4099–4106. ijcai.org.

Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. 2018. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, et al. 2017. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*.

Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, FrÃ©dÃ©ric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales, and Jacopo Staiano. 2020. Project piaf: Building a native french question-answering dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5481–5490, Marseille, France. European Language Resources Association.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32, pages 14014–14024. Curran Associates, Inc.

Sewon Min, Minjoon Seo, and Hannaneh Hajishirzi. 2017. Question answering through transfer learning from large fine-grained supervision data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 510–517, Vancouver, Canada. Association for Computational Linguistics.

Nafise Sadat Moosavi, Angela Fan, Vered Shwartz, Goran Glavaš, Shafiq Joty, Alex Wang, and Thomas Wolf, editors. 2020. *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*. Association for Computational Linguistics, Online.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning {bert}: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. Neural Arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. 2020. Q-BERT: hessian based ultra low precision quantization of BERT. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8815–8821. AAAI Press.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3645–3650. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4003–4012.

Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017. Neural domain adaptation for biomedical question answering. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 281–289, Vancouver, Canada. Association for Computational Linguistics.

Caiming Xiong, Victor Zhong, and Richard Socher. 2017. Dynamic coattention networks for question answering. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Caiming Xiong, Victor Zhong, and Richard Socher. 2018. DCN+: Mixed objective and deep residual coattention for question answering. In *International Conference on Learning Representations*.

Seungyoung Lim;Hyunjeong Lee;Soyoon Park;Myungji Kim Youngmin Kim. 2020. KorQuAD 2.0: Korean QA Dataset for Web Document Machine Comprehension. *Journal of KIISE*, 47:577–586.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2021. Revisiting few-sample {bert} fine-tuning. In *International Conference on Learning Representations*.

Wei Zhang, Lu Hou, Yichun Yin, Lifeng Shang, Xiao Chen, Xin Jiang, and Qun Liu. 2020. TernaryBERT: Distillation-aware ultra-low bit BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 509–521, Online. Association for Computational Linguistics.