Exploring Open Domain Question Answering in French

Espoir Murhabazi Buzina [1]

[1]University of Essex and Masakhane

Abstract

The state-of-the-art model language models have matched human performance in reading for comprehension tasks for both English and French languages.Simultaneously, research has been done on improving those models performance on Open Domain Question Answering (ODQA) in English. However, Open Domain Question Answering (ODQA) in French language has been left behind due to the lack ODQA datasets. In the present work, we built an ODQA dataset from Wikipedia articles and the French Reading Comprehension datasets. Furthermore, we used that dataset to build an end-to-end framework for ODQA in French. The framework used the BM25 algorithm to query the documents and efficiently process them using the fusion in the decoder model as a reader. It achieved an exact match ratio of 58 %, an F1 score of 76 % on the validation set, using only four paragraphs as context. Those results showed that despite using a smaller dataset, we were able to get comparable results to an ODQA model trained on large English datasets.
The code, the dataset and the model used for this work will be made available in the HuggingFace library upon approval of this work.

Hypothesis

The following hypothesis is stated in our work: training the state-of-the-art model for ODQA on a French datasets will yield statistically the same performance as training with the English datasets.

Datasets

The dataset were built from 2 different French Reading for Comprehension datasets: the FQuad [2] and the PIAF [4] dataset.

For each question in the PIAF dataset and FQuad, supporting paragraphs were retrieved from french Wikipedia articles saved in Elasticsearch.

The plot on figure 1 illustrate our dataset size and splits

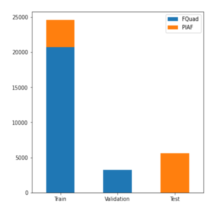

Figure 1. Dataset size for each split

Methods

A standard framework for ODQA consists of two main components, a retriever and a reader. The retriever is a classical information retrieval system that uses the BM25 algorithm to retrieve supporting paragraphs given a question.

A reader is a language model that inputs the question and supporting paragraphs and predicts the answer. The answer prediction can be either reading for comprehension or a text generation task. The plot on the figure 2 illustrates the general framework for ODQA

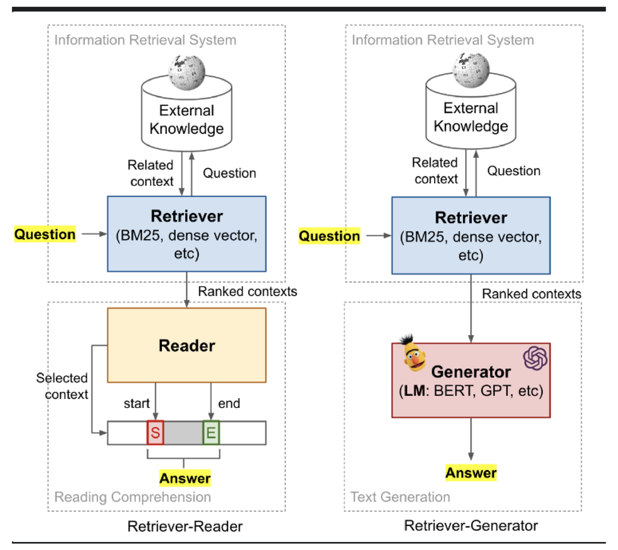

Figure 2. General framework for ODQA. Image source : [6]

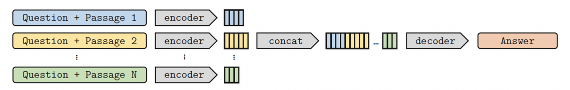

Figure 3. Fusion in decoder approach suggested by [3]

The reader for this work was built on top of the fusion in decoder [3], it is a generative model approach which process multiple paragraphs concurrently in the encoder and produce a vector representation of those paragraphs. Then used that vector in the decoder to generate the answer. Figure 3 illustrates the fusion in the decoder approach.

Key Finding

- Training the state-of-the-art model for ODQA on a smaller French dataset will yield statistically the same performance as training with a large English dataset.
- As for the English, French Language models can leverage paragraphs retrieved from classical information retrieval systems to boost their performances.
- The fusion in the decoder [3] model trained on the french QA dataset with only four contexts paragraphs yields better results for information requests or simple questions such as a question about names, places, and numbers.
- Unfortunately, the model yielded poor results on questions that require long answers because in general they are question that involves reasoning.
- And failed to generalize on a new dataset, datasets that contain paragraphs that were not seen during training.

Evaluation Metrics

Both F1 score and Exact match ratio (EM) are used to evaluate the model's performance for the question-answering model [2]. The EM measures the percentage of prediction that matches the ground truth answer. The F1, on the other hand, computes the average overlap between the prediction and the ground truth tokens.

Results and Analysis

The model yield a f1 score of **76 %** and EM **58.64** on the validation set . And a f1 **57.79** , an EM **41.93** the test set.

The table 1 illustrates how our results compare to those obtained by training the model's same models in English ODQA.

Table 1. Comparison of our method with Results on Squad Open

Dataset	FQuAD Val		SQuAD Val	
	EM	F1	EM	F1
number_context				
Path Retriever [1]	-	-	56.5	63.8
Fusion-in-Decoder (large) [3]	-	-	56.7	63.2
Our method with four context	**58.56**	**76.14**	-	-

The figure 2 illustrates the f1 score and the exact match of our model grouped by question types on our validation set.

Table 2. Results By Question Types

	f1	exact	count	answer_length	number_words
question_type					
qui	0.85	0.76	444	15.98	2.48
quand	0.84	0.68	201	12.73	2.69
combien	0.80	0.64	171	9.93	2.06
quel	0.74	0.57	1029	19.84	3.27
quoi	0.71	0.47	140	28.25	4.58
qu'est	0.68	0.56	50	22.86	3.66
pourquoi	0.68	0.37	63	38.41	6.56

Future Work

While our results are interesting, this work is still in progress, and in future versions we would like to :

- Use neural information retrieval approaches and analyze how the quality of retrieved documents affects the model performance.
- Compress the model into an smaller size and analyze how it perform compared to the baseline approach.

Acknowledgments

We thank the University of Essex for giving us the resources to complete this project.

We warmly thank Dr Jon Chamberlain for his academic guidance and the motivation to complete this paper.

Finally, we would also like to thank the Masakhane community for their mentorship and support in our journey to NLP.

References

[1] Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *International Conference on Learning Representations*, 2020.

[2] Wacim Belblidia, Martin dHoffschmidt, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. Fquad: French question answering dataset. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1193–1208. Association for Computational Linguistics, 2020.

[3] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. pages 874–880.

[4] Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, FrÃ©dÃ©ric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales, and Jacopo Staiano. Project piaf: Building a native french question-answering dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5483–5492, Marseille, France, May 2020. European Language Resources Association.

[5] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[6] Lilian Weng. How to build an open-domain question answering system? 2020.