FrenchMedMCQA: A French Multiple-Choice Question Answering Dataset for Medical domain

Yanis Labrak[1,4] **Adrien Bazoge**[2] **Richard Dufour**[2] **Béatrice Daille**[2]

Pierre-Antoine Gourraud[3] **Emmanuel Morin**[2] **Mickael Rouvier**[1]

LIA - Avignon University[1] LS2N - Nantes University[2]

first.lastname@univ-avignon.fr first.lastname@univ-nantes.fr

CHU de Nantes - La clinique des données - Nantes University[3] Zenidoc [4]

Abstract

This paper introduces FrenchMedMCQA, the first publicly available Multiple-Choice Question Answering (MCQA) dataset in French for medical domain. It is composed of 3,105 questions taken from real exams of the French medical specialization diploma in pharmacy, mixing single and multiple answers. Each instance of the dataset contains an identifier, a question, five possible answers and their manual correction(s). We also propose first baseline models to automatically process this MCQA task in order to report on the current performances and to highlight the difficulty of the task. A detailed analysis of the results showed that it is necessary to have representations adapted to the medical domain or to the MCQA task: in our case, English specialized models yielded better results than generic French ones, even though FrenchMedMCQA is in French. Corpus, models and tools are available online.

1 Introduction

Multiple-Choice Question Answering (MCQA) is a natural language processing (NLP) task that consists in correctly answering a set of questions by selecting one (or more) of the given N candidates answers (also called *options*) while minimizing the number of errors. MCQA is one of the most difficult NLP tasks because it requires more advanced reading comprehension skills and external sources of knowledge to reach decent performance.

In MCQA, we can distinguish two types of answers: (1) single and (2) multiple ones. Most datasets focus on single answer questions, such as MCTest (Richardson et al., 2013), ARC-challenge (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), QASC (Khot et al., 2019), Social-IQA (Sap et al., 2019), or RACE (Lai et al., 2017). To our knowledge, few studies have been done to construct medical MCQA dataset. We can cite the MedMCQA (Pal et al., 2022) and HEAD-QA (Vilares and Gómez-Rodríguez, 2019) corpora which contain single answer questions in Spanish and English respectively. For the multiple answer questions, MLEC-QA (Li et al., 2021) provides 136k questions in Chinese covering various biomedical sub-fields, such as clinic, public health and traditional Chinese medicine.

The French community has recently greatly increased its efforts to collect and distribute medical corpora. Even if no open language model is currently available, we can cite the named entity recognition (Névéol et al., 2014) and information extraction (Grabar et al., 2018) tasks. However, they remain relatively classic, current approaches already reaching a high level of performance.

In this article, we introduce FrenchMedMCQA, the first publicly available MCQA corpus in French related to the medical field, and more particularly in the pharmacological domain. This dataset contains questions taken from real exams of the French diploma in pharmacy. Among the difficulties related to the task, the questions asked may require a single answer for some and multiple ones for others. We also propose to evaluate state-of-the-art MCQA approaches, including an original evaluation of several word representations across languages.

Main contributions of the paper concern (1) the distribution of an original MCQA dataset in French related to the medical field, (2) a state-of-the-art approach on this task and a first analysis of the results, and (3) an open corpus, including tools and models, all available online.

2 The FrenchMedMCQA Dataset

In this section, we detail the FrenchMedMCQA dataset and discuss data collection and distribution.

2.1 Dataset collection

The questions and their associated candidate answer(s) were collected from real French pharmacy

exams on the remede[1] website. This site was built around a community linked to the medical field (medicine, pharmacy, odontology...), offering multiple information (news, job offers, forums...) both for students and also professionals in these sectors of activity. Questions and answers were manually created by medical experts and used during examinations. The dataset is composed of 2,025 questions with multiple answers and 1,080 with a single one, for a total of 3,105 questions. Each instance of the dataset contains an identifier, a question, five options (labeled from `A` to `E`) and correct answer(s). The average question length is 14.17 tokens and the average answer length is 6.44 tokens. The vocabulary size is of 13k words, of which 3.8k are estimated medical domain-specific words (*i.e.* related to the medical field). We find an average of 2.5 medical domain-specific words in each question (17% of words in average of a question) and 2.0 in each answer (36% of words in average of an answer). On average, a targeted medical domain-specific word is present in 2 questions and in 8 answers.

2.2 Dataset distribution

Table 1 presents the proposed FrenchMedMCQA dataset distribution for the train, development (dev) and test sets detailed per number of answers (*i.e.* number of correct responses per question). Globally, 70% of the questions are kept for the train, 10% for validation and last 20% for testing.

# Answers	Training	Validation	Test	Total
1	595	164	321	**1,080**
2	528	45	97	**670**
3	718	71	141	**930**
4	296	30	56	**382**
5	34	2	7	**43**
Total	**2171**	**312**	**622**	**3,105**

Table 1: FrenchMedMCQA dataset distribution.

3 Methods

The use alone of the question to automatically find the right answer(s) is not sufficient in the context of a MCQA task. State-of-the-art approaches then require external knowledge to improve system performances (Izacard and Grave, 2020; Khashabi et al., 2020). In our case, we decide to build a two-step retriever-reader architecture comparable to UnifiedQA (Khashabi et al., 2020), where the retriever job is to extract knowledge from an external corpus and using it by the reader to predict the correct answers for each question. Figure 1 presents the two-step general pipeline, first step being the retriever module, that extracts external context from the question (see Section 3.1), and second step being the reader, called here question-answering module (see Section 3.2), that automatically selects answer(s) to the targeted question.

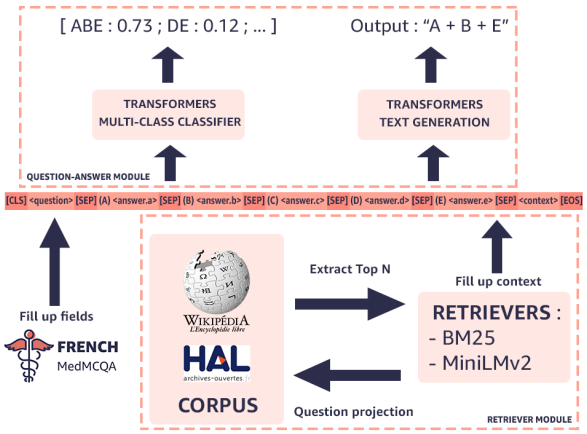

Figure 1: Steps of the pipeline.

3.1 Retriever module

An external medical-related corpus fully composed of French has first been collected from two online sources: Wikipedia life science and HAL, the latter being an open archive run by the French National Centre for Scientific Research (CNRS) where authors can deposit scholarly documents from all academic fields. In our case, we focus on extracting papers and thesis from various specialization, such as Human health and pathology, Cancerology, Public health and epidemiology, Immunology, Pharmaceutical sciences, Psychiatric disorders and Drugs. This results in 1 million of passages (*i.e.* a portion of text that contains at least 100 characters) in HAL and 286k passages in Wikipedia.

This corpus is then used as a context extension for a question. We therefore used a retriever pipeline to automatically assign questions to the most likely passage in the external source. Two retrieval approaches are compared in this article:

- BM25 Okapi (Trotman et al., 2014) for the implementation of the base BM25 algorithm (Robertson and Sparck Jones, 1988).
- SentenceTransformers framework (Reimers

[1] http://www.remede.org/internat/pharmacie/qcm-internat.html

and Gurevych, 2019) is used to perform semantic search using state-of-the-art language representations taken from Huggingface's Transformers library (Wolf et al., 2019).

For both approaches, the goal is to embed each passage of the external corpus into a vector space using one of the two representations. On its side, the question is concatenated with the five options (*i.e.* answers associated to the question) to form a new query embedded in the same vector space. Embeddings from question and passages are finally compared to return the closest passages of a query (here, the cosine similarity is the distance metric). For the SentenceTransformers approach, we used a fast and non domain specific model called `MiniLMv2` (Wang et al., 2020). Note that the 1-best passage is only used in these experiments.

3.2 Question-answer module

A goal of our experiments was to compare baseline approaches regarding two different paradigms. The first one is referred to a discriminative approach and consists in assigning one of N classes to the input based on their projection in a multidimensional space. We also referred to it as a multi-class task. At the opposite, the second method is a generative one which consists of generating a sequence of tokens, also called *free text*, based on a sequence of input tokens identical to the one used for the discriminative approach. The difference with the discriminative approach lies in the fact that we are not outputting a single class, like `ABE` for the question `6234176387997480960`, but a sequence of tokens following the rules of the natural language and referring to a combination of classes like `A + B + E` in the case of our studied generative model (see Section 3.2.2).

3.2.1 Discriminative representations

Four discriminative representations are studied in this paper. We firstly propose to use **CamemBERT** (Martin et al., 2020), a generic French pre-trained language model based on RoBERTa (Liu et al., 2019). Since no language representation adapted to the medical domain are publicly available for French, we propose to evaluate the two pre-trained representations **BioBERT** (Lee et al., 2019) and **PubMedBERT** (Gu et al., 2022), both trained on English medical data and reaching SOTA results on biomedical NLP tasks, including QA (Pal et al., 2022). Finally, we consider a multilingual generic pre-trained model, **XLM-RoBERTa** (Conneau et al., 2020) based on RoBERTa, to evaluate the gap in terms of performance with CamemBERT.

3.2.2 Generative representation

Recently, generative models have demonstrated their interest on several NLP tasks, in particular for text generation and comprehension tasks. Among these approaches, **BART** (Lewis et al., 2019) is a denoising autoencoder built with a sequence-to-sequence model. Due to its bidirectional encoder and left-to-right decoder, it can be considered as generalizing BERT and GPT (Radford et al., 2019), respectively. BART training has two stages: (1) a noising function used to corrupt the input text, and (2) a sequence-to-sequence model learned to reconstruct the original input text. We then propose to evaluate this representation in this paper.

4 Experimental protocol

Each studied discriminative and generative model is fine-tuned on the MCQA task with FrenchMedMCQA training data using an input sequence composed of a question, its associated options (*i.e.* possible answers) and its additional context, all separated with a "[SEP]" token, e.g. [CLS] <question> [SEP] (A) <answer.a> [SEP] (B) <answer.b> [SEP] (C) <answer.c> [SEP] (D) <answer.d> [SEP] (E) <answer.e> [SEP] <context> [EOS].

For each question, the context is the text passage with highest confidence rate and can either be obtained using the BM25 algorithm or semantic search as described in Section 3.1.

Concerning the outputs of the systems, we have for the BART generative model a plain text containing the letter of the answers from A to E separated with plus signs in case of the questions with multiple answers, *e.g.* A + D + E. For the other architectures (*i.e.* discriminative approaches), we simplify the multi-label problem into a multi-class one by classifying the inputs into one of the 31 existing combinations in the corpus. Here, a class may be a combination of multiple labels, *e.g.* if the correct answers are the A and B ones, then we consider the correct class being AB, which explains the number of 31 classes.

4.1 Evaluation metrics

The majority of tasks concentrate either on multi-class or binary classification since they have a single class at a time. However, occasionally, we will

	Without Context		Wiki w/ BM25		HAL w/ BM25		Wiki w/ MiniLMv2		HAL w/ MiniLMv2	
Architecture	Hamming	EMR	Hamming	EMR	Hamming	EMR	Hamming	EMR	Hamming	EMR
BioBERT V1.1	36.19	15.43	**38.72**	16.72	33.33	14.14	35.13	16.23	34.27	13.98
PubMedBERT	33.98	14.14	34.00	13.98	35.66	15.59	33.87	14.79	35.44	14.79
CamemBERT-base	36.24	16.55	34.19	14.46	34.78	15.43	34.66	14.79	34.61	14.95
XLM-RoBERTa-base	37.92	17.20	31.26	11.89	35.84	16.07	32.47	14.63	33.00	14.95
BART-base	31.93	15.91	34.98	**18.64**	33.80	17.68	29.65	12.86	34.65	18.32

Table 2: Performance (in %) on the test set using the Hamming score and EMR metrics.

have a task where each observation has many labels. In this case, we would have different metrics to evaluate the system itself because multi-label prediction has an additional notion of being partially correct. Here, we focused on two metrics called the Hamming score (commonly also multi-label accuracy) and Exact Match Ratio (EMR).

4.1.1 Hamming score

The accuracy for each instance is defined as the proportion of the predicted correct labels to the total number (predicted and actual) of labels for that instance. Overall accuracy is the average across all instances. It is less ambiguously referred to as the Hamming score rather than Multi-label Accuracy.

4.1.2 Exact Match Ratio (EMR)

The Exact Match Ratio (EMR) is the percentage of predictions matching exactly the ground truth answers. To be computed, we sum the number of fully correct questions divided by the total number of questions available in the set. A question is considered *fully correct* when the predictions are exactly equal to the ground truth answers for the question (*e.g.* all multiple answers should be correct to count as a correct question).

5 Results

Table 2 compiled the performance (in terms of Hamming score and EMR) of all the studied architectures and retrievers pipelines. For sake of comparison, the column *Without Context* has been added, considering that no retriever is used (*i.e.* no external passage is present in the QA system).

As we can see, the best performing model is different according to the used metric. **BioBERT V1.1** reaches best performance using the Hamming score and **BART-base** in the case of the EMR. These first observations are quite surprising since both models are trained on English data. While we could expect higher performance with French models (CamemBERT for example), the fact that these models are trained on specialized data for one (BioBERT) and on a model designed for the targeted task (SOTA on question-answering for BART) finally shows that language models trained on generic data are inefficient for the MCQA task on medical domain.

In all considered architectures, context seems to have a small impact on systems performance, with a limited increase or drop depending on the configurations. Clearly, the **RoBERTa** performance is much higher without context (*i.e.* without the use of the retriever part), while models based on **BERT** generally (8 times on 12) outperform their own baseline performances with external context. The fact that we consider the 1-best passage only may explain this impact.

Concerning **XLM-RoBERTa-base** (cross lingual representation), we obtain in the case of the context extracted using BM25 from Wikipedia, the worst Hamming score and EMR out of all the discriminative approaches. This confirms our first observation that a non-specialized model does not allow to achieve the best performance on this task.

Using BM25 promotes better context than semantic search using **MiniLMv2** on both Wikipedia and HAL for most of the runs. Finally, the source depends of the retriever and model used. A majority of the experiments demonstrate that HAL outperforms Wikipedia on BM25 despite the fact that the best model was obtained using Wikipedia.

The scripts to replicate the experiments[2] as well as the pre-trained models[3] are available online.

6 Conclusion

We proposed in this paper FrenchMedMCQA, an original, open and publicly available Multiple-Choice Question Answering (MCQA) dataset in the medical field. This is the first French corpus in this domain, including single and multiple answers to questions. Several state-the art systems have been evaluated to show current performance on the dataset. The analysis of these first results notably

[2] https://github.com/qanastek/FrenchMedMCQA

[3] https://huggingface.co/qanastek/FrenchMedMCQA-BioBERT-V1.1-Wikipedia-BM25/tree/main

highlighted the fact that language models specialized to the medical domain allow us to reach better performance than generic models, even if these have been trained in a different language (here, English biomedical models applied to French).

In future works, we will focus on improving the existing methods for the task of MCQA, considering other strategies for the retriever module (multiple passages, combining contexts...). Likewise, we will also consider the construction of data representation models for French specialized for medical domain.

7 Acknowledgments

This work was financially supported by Zenidoc, the DIETS project financed by the Agence Nationale de la Recherche (ANR) under contract ANR-20-CE23-0005 and the ANR AIBy4 (ANR-20-THIA-0011). This work was performed using HPC resources from GENCI-IDRIS (Grant 2022-AD011013061R1 and 2022-AD011013715).

References

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

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.

Natalia Grabar, Vincent Claveau, and Clément Dalloux. 2018. CAS: French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 122–128, Brussels, Belgium. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2019. Qasc: A dataset for question answering via sentence composition.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Jing Li, Shangping Zhong, and Kaizhi Chen. 2021. MLEC-QA: A Chinese Multi-Choice Biomedical Question Answering Dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8862–8874, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

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.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset, and Pierre Zweigenbaum. 2014. The QUAERO French medical corpus: A ressource for medical entity recognition and normalization. In *Proc of BioTextMining Work*, pages 24–30.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

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

Stephen E. Robertson and Karen Sparck Jones. 1988. *Relevance Weighting of Search Terms*, page 143–160. Taylor Graham Publishing, GBR.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*, ADCS '14, page 58–65, New York, NY, USA. Association for Computing Machinery.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2020. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.