



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

Yanis Labrak, Adrien Bazoge, Richard Dufour, Béatrice Daille, Pierre-antoine Gourraud, Emmanuel Morin, Mickaël Rouvier

► To cite this version:

Yanis Labrak, Adrien Bazoge, Richard Dufour, Béatrice Daille, Pierre-antoine Gourraud, et al.. FrenchMedMCQA: A French Multiple-Choice Question Answering Dataset for Medical domain. Empirical Methods in Natural Language Processing 2022, Dec 2022, Abu Dhabi, United Arab Emirates. , 2022. hal-03913329

HAL Id: hal-03913329

<https://hal.science/hal-03913329>

Submitted on 30 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

Summary

Context

- Multiple Choice Question Answering (MCQA) Task: Selection of the correct candidates (one or more) regarding a question
- Few resources available in the biomedical field
- Data mainly in English and Chinese

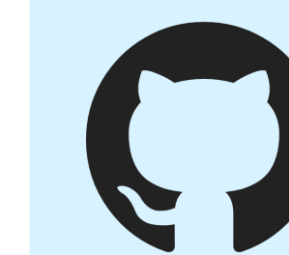
Difficulties

- MCQA is one of the most difficult NLP tasks
- Requires advanced reading comprehension skills and external sources of knowledge
- Freely available textual data are rare for biomedical domain

Contributions

- SOTA models for biomedical multiple-choice question answering in French
- An external passage retriever from various biomedical textual sources
- An open corpus, including tools and models, all available online

Corpus and Tools distribution



github.com/qanastek/FrenchMedMCQA



FrenchMedMCQA-BART-base-Wikipedia-BM25
FrenchMedMCQA-BioBERT-V1.1-Wikipedia-BM25



qanastek/FrenchMedMCQA



Corpus creation and Evaluation

- Collected from real French pharmacy exams (remede.org)
- Questions and answers manually created by medical experts and used during examinations
- Contains an identifier, a question, five options and correct answer(s)

# 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	2,171	312	622	3,105

Multi-label tasks need different metrics, since an observation can be partially correct:

Exact Match Ratio (EMR) $\rightarrow \frac{\text{Number of fully correct questions}}{\text{Total number of questions}}$

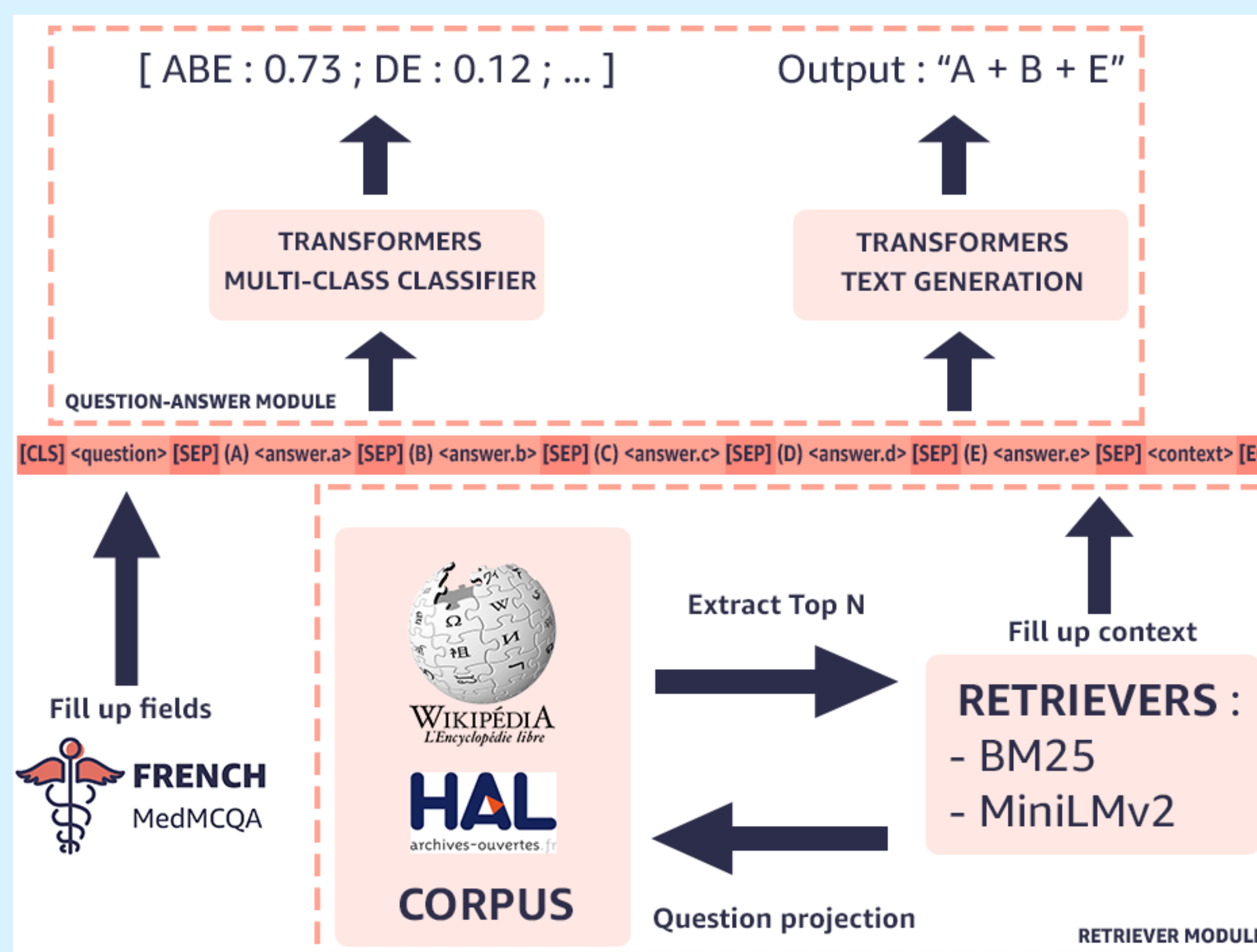
Hamming Score $\rightarrow \frac{1}{N} \sum_{I \in \text{Corpus}} \frac{I_{\text{Truth}}}{I_{\text{Pred}} + I_{\text{Ref}}}$

State of the art MCQA system

Discriminative representations:

- CamemBERT, a generic French pre-trained language model based on RoBERTa
- Medical domain pre-trained representations in English BioBERT and PubMedBERT

Generative representation: Prompting with BART



External medical-related corpus in French from two online sources:

- Wikipedia life science in French
- HAL papers and thesis from various specializations, such as:
 - Human health and pathology
 - Cancerology
 - Public health and epidemiology
 - Immunology
 - Pharmaceutical sciences
 - Psychiatric disorders
 - Drugs

1 million passages (at least 100 characters) from HAL and 286K from Wikipedia.

	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

Conclusion and Perspectives

- BioBERT V1.1 reaches best performance on Hamming score and BART-base for EMR
- Best performing models are based on domain specific English words representations while being applied on French
- SOTA domain specific models outperform generic language specific model
- Future work: propose an open French language model specialized on biomedical domain, still doesn't exist

Acknowledgments

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)

