Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 14, Issue. 1 : 2023
ISSN : **1906-9685**

A New Era of French Biomedical NLP: The FrBMedQA Dataset

Mohapatra Girashree Sahu,Nibedita Sial,Saroj Nanda
Dept. of Computer Science and Engineering,Gandhi Institute For Technology,Bhubaneswar,752054
Email: girashree@gift.edu.in

ABSTRACT

FrBMedQA is the first French biomedical question answering dataset, containing 41k+ passage-question instances. It was automatically constructed in a cloze-style manner, from biomedical French Wikipedia articles. To test the validity and difficulty of the dataset, we experimented with four statistical baseline models, a biomedical bidirectional encoder representation from transformers (BERT)-based model, and two French BERT-based language model. We also did human evaluation on a subset of the test set. All the three tested models were not able to surpass the best performing baseline model. Human performance at 61.11% is leading the leaderboard with more than 8% from the best performing model. We made available the dataset and the code to reproduce our results.

1. INTRODUCTION

Question answering (QA) is the task of extracting or generating a valid answer to a question from a passage, a document, or a set of documents. General QA has seen huge progress in recent years, largely due to the numerous datasets [1]–[5] that has been released. While there is no shortage of English QA datasets, there is only two French ones [6], [7].

Biomedical question answering (BQA) can be considered as a sub-task of general QA, which is concerned with finding relevant answers to questions in biomedical text. In contrast with general QA, there is a limited number of BQA datasets [8]. Adding to that, the majority of these datasets are very small in size. The text retrieval conference (TREC) Genomics Track [9] for example, has only 28 questions. BioASQ [10] also has only 3k question-answer instances. The other datasets are ether automatically constructed or artificially generated. Building a human annotated BQA dataset is costly and time-consuming, as the annotation must be done by medical experts. But, these datasets are of more quality than the automatically constructed or artificially generated ones that suffers from a high noise ratio. Nevertheless, applying statistical or deep learning models to the task of BQA requires a large number of training samples, therefore, the large automatically constructed and artificially generated datasets are equally helpful as the small size human annotated ones.

The main technique used to automatically construct a QA dataset is the cloze-style [11] technique. Which translates the QA task into a fill in the blank problem, by hiding a word or a set of words from a text, asking to guess the hidden tokens by going back to the context text of which they were taking. The first dataset that adopted this technique in the QA domain was the children's books dataset [12]. The well-known cable news network (CNN) and Daily Mail datasets [2] used the same technique.

Several BQA datasets also used the same cloze-style technique. The first one being BioRead [13], followed by biomedical knowledge comprehension (BMKC) [14], and biomedical machine reading comprehension (BioMRC) [15]. In the context of BQA, the hidden word must be a biomedical term, so these datasets used biomedical annotation tools to automatically identity biomedical terms.

While there are numerous English BQA datasets, there is no French BQA dataset at the moment of writing. This is arguably the number one challenge for French BQA, as datasets are the first prerequisite for training and evaluating systems. As a first step toward solving this challenge, we introduce the FrBMedQA dataset. With more than 41k instances, the dataset was collected from Wikipedia French biomedical articles, and then constructed in a cloze-style manner similar to the other mentioned BQA datasets.

To evaluate the validity and difficulty of the dataset, also as a first step towards a public leaderboard, we implemented and experimented with several baseline models, a neural-based biomedical language model, and two French monolingual language models. We also did human evaluations of a subset of the test set. We made available the dataset and the code to reproduce our results.

We organised the rest of this paper in the following way, in the next section we overview the related work that has been done in BQA datasets, and in French QA. In section three, we describe in details the following aspects of the FrBMedQA dataset, corpus retrieval and annotation, the cloze-style strategy we used, and we give a detailed analysis of the textual and biomedical properties of the dataset. In section four, we describe the baseline models as well as the neural-based ones that we applied to the dataset, also giving the results of our experiments and discussion them. We finish the paper with a conclusion where we also talk about future research directions following the public release of the dataset.

2. RELATED WORK

BQA datasets: While there is no French BQA dataset to date, numerous ones exists for English. In 2006, The TREC Genomics Track [9] included a QA task for the first time. A dataset was constructed for this task, by collecting more than 162k full-text biomedical research articles. Participating systems were required to retrieve relevant passages from the research articles in response to a question, these passages were then evaluated by human judges. While this dataset was a pioneer in BQA, it suffered from two big limitations. Only 28 questions were provided along with the full-text articles. More importantly, no question-relevant passages instances were provided for training. Therefore, this dataset could not be used by machine learning (ML) or deep learning (DL) systems.

The most well-known dataset in BQA is BioASQ [10], unlike the majority of BQA datasets, BioASQ is manually annotated by medical experts, although it only contains ~3k instances. This dataset was first released in 2015 as part of a larger long-running biomedical natural language processing (NLP) competition that takes place every year. Questions in this dataset are of four types: factoid, yes/no, list, and summary. The only limiting factor of this dataset is its small size.

In response to the problem of small size of BQA datasets, BioRead [13] was introduced in 2018. It is currently the largest BQA dataset with ~16.4 million instances. It was automatically constructed from biomedical PubMed full-text research articles following the cloze-style technique. A predefined number of sentences is selected each time as a passage, with the following sentence as the question. MetaMap [16] was used to annotate the biomedical entities found in the passage and question, an entity from the question is then masked. The task then is to find the masked entity from all the entities in the passage as possible candidates. With the obvious advantage of this dataset being its size, it suffers from two limitations, while for example BioASQ supports four types of questions, BioRead only supports one type being one choice selection as a result of the adopted cloze-style technique. The other limitation is the fact that many passage-question instances were taking from the references section, figure and table captions, footnotes, etc.

BMKC [14] is another cloze-style BQA dataset with ~500k passage-question instances, constructed from abstracts of PubMed biomedical research papers. The title of the articles was chosen as the question, in another setting they used the last sentence of the abstract as the question. The choice to only use abstracts and titles was taking to reduce the noise. Like BioRead, the authors of this dataset automatically annotated biomedical entities in the question and the passage, and masked an entity from the question to be guessed then from the other entities in the passage.

Another automatically annotated dataset is MedQuAD [17] containing more than 47k question-answer pairs extracted and generated from different trusted medical websites. PubMedQA [18] that is constructed using PubMed abstracts and has 1k expert-annotated, ~61k unlabeled and ~211k artificially generated QA instances. This dataset only supports yes/no/maybe questions.

In 2020, the BioMRC dataset [15] was introduced as an improved version of BioRead. It has 812k passage-question instances, and follows the same cloze-style strategy as its predecessor. To reduce noise, which is the main drawback of BioRead, the authors of BioMRC only used abstracts to generate passage-question instances, in contrast to using the full-text in BioRead. They also used disease named entity

recognition and normalization (DNORM) biomedical entity annotations [19] which is more accurate than MetaMap that was used in BioRead.

French QA: To the best of our knowledge, there is currently only two French QA datasets, and no BQA ones. FQuAD [6] was the first French QA dataset that was introduced in 2020. It contains ~60k manually annotated passage-question instances. Same as SQuAD, data was collected from Wikipedia following the same strategy. To experiment with the dataset, the authors applied two families of models, native French monolingual models using CamemBERT [20] and FlauBERT [21], and multilingual models with mBERT [22], and XLM-RoBERTa [23]. Experiments showed that the native French monolingual CamemBERT model performed better.

Shortly after the release of FQuAD, the PIAF [7] dataset was introduced. Similar to FQuAD, it was collected from Wikipedia following the same strategy. But in contrast to FQuAD, it only contains 3835 question-answer pairs. To experiment with the dataset, the authors of the dataset only tested the best performing model on FQuAD, which was CamemBERT using multiple fine-tuning strategies.

3. METHOD

3.1. Corpus retrieval and annotation

Most BQA datasets use PubMed as the source of their corpus. In our case, we could not use it, because PubMed does not provide abstracts for articles written in French. On the other hand, French PubMed equivalents like Lissa prohibits the use or redistribution of abstracts. Faced with this challenge, we decided to use Wikipedia as the source of our dataset.

Out from the five million French articles on Wikipedia, first we retrieved 243k biomedical articles, using two filtering strategies. For the first one, we retrieved every article having a biomedical InfoBox . The second one was, retrieving articles having at least one biomedical term in their title, and at least ten biomedical terms in their text. To do that, we relied on a list of biomedical terms that we constructed from various French biomedical dictionaries. After the filtering, we cleaned up the text from noisy parts like references and some html tags, and split it into paragraphs. We discarded paragraphs composed of less than three sentences, or containing less than 23 tokens. After collecting the corpus, we used Semantic Indexing of French biomedical Resources (SIFR) annotator [24], a French biomedical named entity recognition (NER) tool, to annotate the biomedical entities found in the corpus. To limit the annotation strictly to the biomedical terms, we only considered entities belonging to the unified medical language system (UMLS) semantic groups shown in Table 1.

Table 1. UMLS semantic groups considered for annotation

Semantic group	ID	Number of entities	Percentage
Chemicals & drugs	CHEM	62,820	29.77%
Anatomy	ANAT	54,906	26.02%
Physiology	PHYS	30,660	14.53%
Disorders	DISO	30,045	14.24%
Phenomena	PHEN	16,007	7.59%
Procedures	PROC	12,588	5.96%
Genes & molecular sequences	GENE	3,475	1.75%
Devices	DEVI	476	0.02%

3.2. Cloze-style instance generation technique

An instance of the FrBMedQA dataset is a tuple containing, a context passage, a question, candidate answers, and the answer. The task then, is to select the correct answer for the question from the list of candidate answers also appearing in the passage. To generate the instances we used the same cloze-style strategy used to construct the CNN and daily mail datasets, with some minor modifications. First we replaced all biomedical entities with pseudo-tokens of the form @entityID, where ID is a unique integer identifier for each biomedical entity, we start with ID zero and increment it with each new biomedical entity. Other cloze-style BQA datasets like BioRead, and BioMRC, follow two strategies when replacing biomedical entities with pseudo-tokens. The first one being, restarting from ID zero for each new instance, and the second one being, maintaining the same ID for the same biomedical entity for all instances. We chose to only follow the second strategy after seeing that it gives the best results for neural-based systems on BioRead, and BioMRC. This is because, in the second setting, neural-based systems are able to learn useful properties of pseudo-tokens from training on multiple instances. Anonymizing biomedical terms by replacing them with pseudo-token prevent systems from using background knowledge, and to force them to read and comprehend the context passage. Without the anonymization, even an n-gram model previously trained on the same corpora will be able to retrieve the missing @placeholder.

To generate the passage and the question, we go through all sentences in a paragraph starting from the first one, we then search for pseudo-tokens in the current sentence, that are also appearing in the rest of the sentences. When we find a match, we choose the current sentence as the question, and the rest of the sentences as the passage. This approach is different from the other cloze-style datasets, where the question is ether the first or the last sentence of the paragraph. With our approach, we were able to have more instances, as in a lot of cases the first or the last sentence don't share biomedical entities with the rest of the text. However, we make sure not to have a sentence chosen as a question in other passages. After that, we replace the pseudo-token that we found in the question with a placeholder of the form @placeholder, the pseudo-token then became the answer, and the set of all pseudo-tokens in the context and question became the candidate answers. If multiple pseudo-tokens are found in the question, the same operation is repeated for every pseudo-token. To further illustrate the process of corpus collection, annotation, and instance generation, algorithm 1 shows the exact algorithm that we used. An example of a random instance from the dataset is shown in Figure 1, showing the context, question, candidate entities, and the answer, before and after the application of the cloze-style encoding step.

Algorithm 1: The overall algorithm of corpus collection, annotation, and instance generation

```
Input:
    data ← French Wikipedia data
    infoBoxes ← list of Wikipedia biomedical InfoBoxes
    medTerms ← list of French biomedical terms
    semanticGroups ← list of allowed semantic groups
Output: List of instances as (passage, question, candidate answers, answer) tuples
articles ← {}
For each article in data do
  if article.InfoBox in infoBoxes or article.title has one of medTerms then
    Remove noisy tokens from article
    paragraphs ← split article into paragraphs
    for each paragraph in paragraphs do
      annotations ← call the SIFR annotator web service with paragraph text
      for each entity in annotations do
        if entity.semanticGroup in semanticGroups then
          pseudoToken ← generate pseudo-token replacement for entity
          replace each occurrence of entity in paragraph by pseudoToken
          sentences ← split paragraph into sentences
          for each sentence in sentences do
            restOfText ← sentences - sentence
            if sentence has pseudoToken and restOfTex has pseudoToken then
              question ← replace the pseudoToken in sentence by @placeholder
              candidateAnswers ← list of all pseudo-tokens in paragraph
              articles ← append: {restOfText, question, candidateAnswers, pseudoToken}
            end if
          end for
        end if
      end for
    end for
  end if
end for
```

3.3. Dataset analysis

The dataset is divided into three sets, 80% as training set, 10% for validation, and the remaining 10% as the test set. Table 2 shows different statistics about the dataset, like the number of instances in each set. Also, the average, maximum, and minimum length of the question, context, and the candidate answers.

Table 2. Dataset statistics (length in tokens)

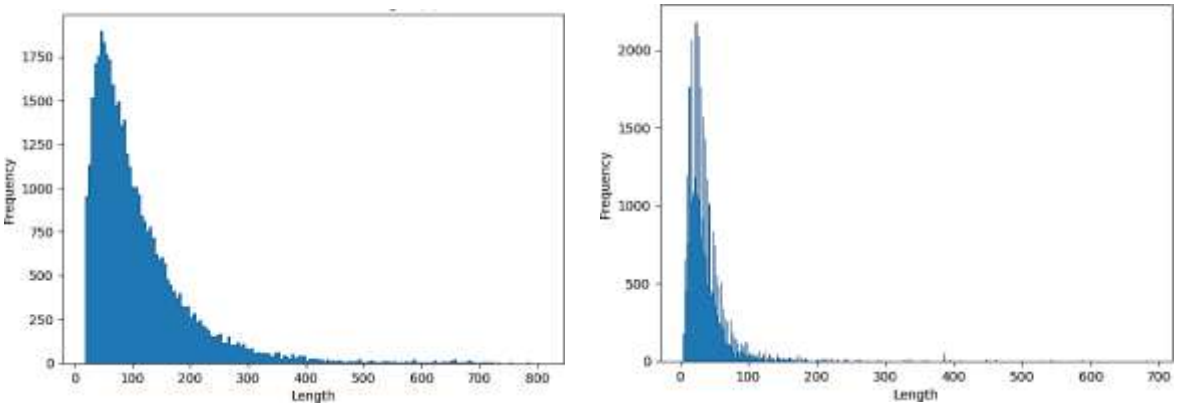
	Training	Validation	Test	Total
Instances	32,888	4,111	4,110	41,109
Avg # candidates	4.81	4.89	4.73	4.81
Max # candidates	42	38	28	42
Min # candidates	2	2	2	2
Avg context len.	111.02	111.44	109.94	110.95
Max context len.	807	715	715	807
Min context len.	19	19	19	19
Avg question len.	38.54	39	38.32	38.56
Max question len.	685	587	542	685
Min question len.	5	5	5	5

Passage: Le traitement secondaire va être le traitement définitif. L'enfant est mis sous traitement antibiotique avant l'opération et il le continuera 48 heures après qu'elle sera passée. Ce traitement permet de prévenir toute infection, car c'est la plus grosse complication possible. Elle peut être due à l'entérocolite ou à une contamination par les selles. Cette chirurgie peut être pratiquée avant les 3 premiers mois de vie si nécessaire. Il existe différentes techniques pour ce traitement mais toutes ces opérations ont le même but. Elles permettent de rétablir la continuité du tube digestif après avoir effectué une ablation partielle de la partie du côlon malade. Après cette ablation de la portion pathologique du côlon, le segment de l'iléon est relié au segment du côlon qui reste, avec du fil ou des agrafes. Cette intervention n'entraîne généralement pas de conséquences sur le fonctionnement du tube digestif. Quand l'ensemble du côlon est atteint, c'est l'iléon normalement innervé qui doit être amené au niveau du rectum ou même de l'anus. Une technique alternative consiste en une résection transanale du colon distal avec des résultats comparables voire supérieurs à la technique classique	**Passage**: Le traitement secondaire va être le traitement définitif. L'enfant est mis sous traitement antibiotique avant l'opération et il le continuera 48 heures après qu'elle sera passée. Ce traitement permet de prévenir toute @entity0, car c'est la plus grosse complication possible. Elle peut être due à l'@entity1 ou à une contamination par les selles. Cette chirurgie peut être pratiquée avant les 3 premiers mois de vie si nécessaire. Il existe différentes techniques pour ce traitement mais toutes ces opérations ont le même but. Elles permettent de rétablir la continuité du @entity2 après avoir effectué une ablation partielle de la partie du @entity3 malade. Après cette ablation de la portion pathologique du @entity3, le segment de l'iléon est relié au segment du @entity3 qui reste, avec du fil ou des agrafes. Cette intervention n'entraîne généralement pas de conséquences sur le fonctionnement du @entity2. Quand l'ensemble du @entity3 est atteint, c'est l'iléon normalement innervé qui doit être amené au niveau du @entity4 ou même de l'anus. Une technique alternative consiste en une résection transanale du colon distal avec des résultats comparables voire supérieurs à la technique classique
Question: Autrement dit, le but recherché est de supprimer les zones intestinales ne contenant plus de cellules neuro-ganglionnaires, et de relier les intestins qui fonctionnent normalement à la partie terminale du tube digestif, c'est-à-dire le rectum, si celui-ci possède ces cellules, sinon à l'anus	**Question**: Autrement dit, le but recherché est de supprimer les zones intestinales ne contenant plus de @entity5 neuro-ganglionnaires, et de relier les @entity6 qui fonctionnent normalement à la partie terminale du @entity2, c'est-à-dire le @placeholder, si celui-ci possède ces @entity5, sinon à l'anus
Candidates: Infection, Entérocolite, Tube digestif, Côlon, Rectum, Cellules, Intestins	**Candidates**: @entity0: Infection @entity1: Entérocolite @entity2: Tube digestif @entity3: Côlon @entity4: Rectum @entity5: Cellules @entity6: Intestins
Answer: Rectum	**Answer**: @entity4: Rectum

Figure 1. Sample instance showing the context, question, candidate entities, and the answer, before and after the application of the cloze-style encoding step

In the next update of the dataset, we plan to increase the size from the current 41k+ instances. We also plan to decrease the difference between the maximum and the minimum number of candidate answers. Study the correlation between performance results and passage and question lengths, so we can choose the most adequate maximum and minimum length boundaries. Doing these optimizations on the current state of the dataset will result in a sharp decrease of its size. We plan to gather more French biomedical data to be able to discard instances having some specific properties.

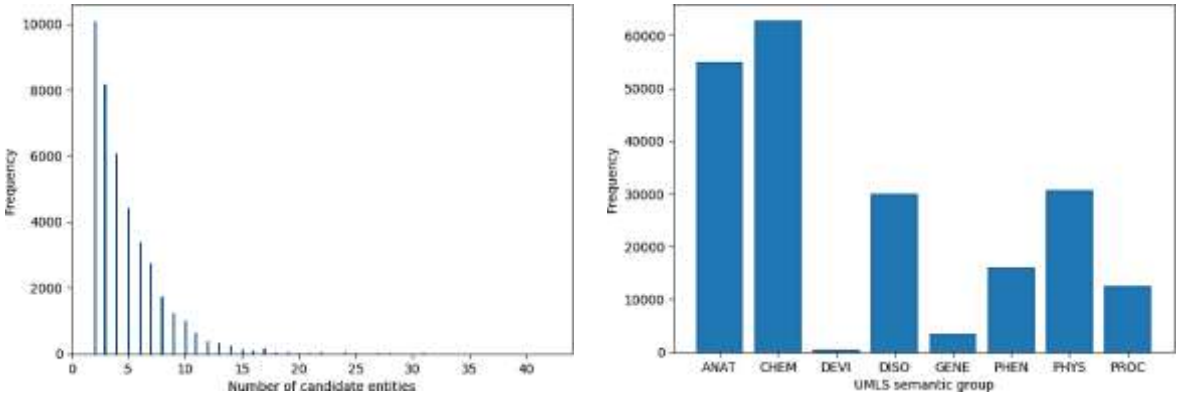
Figure 2 shows the distribution of the context (passage) length in Figure 2(a), and of the question length in Figure 2(b). The majority of instances are grouped around the mean value of 110.95 for context length, and of 38.56 for question length. Having more data in the future will help us discard the outliers.



(a) (b)

Figure 2. The distribution of (a) the context length, and (b) question length

Figure 3 shows the distribution of the number of candidate entities in Figure 3(a), and the distribution of biomedical entities by UMLS semantic group in Figure 3(b). The majority of instances have between two and four candidate answer entities. A system that is randomly guessing an entity as an answer from the list of candidate entities, can actually perform well. To overcome this limitation, we intend to discard instances having less than four unique entities in the next update of the dataset. As for the distribution of the biomedical entities by UMLS semantic group, we think that it does correctly reflect the reality of biomedical textual corpora.



(a) (b)

Figure 3. The distribution of (a) the number of candidate entities, and (b) biomedical entities by UMLS semantic group

4. RESULTS AND DISCUSSION

In order to test the validity and the difficulty of the dataset, and also to offer a first step toward a leaderboard for the dataset, we implemented and experimented with three baseline models already tested on BioMRC plus another baseline model. We also implemented and tested three neural BERT-based models. In addition to that, we did human evaluation on a sub-set of the test set. The following are the models we experimented with.

- Baseline 1: randomly selecting an entity from the list of candidate entities. The table of experiment results shows the mean of three runs.
- Baseline 2: returns the entity (@entityID) that occurs most in the passage and the question, on the ground that this entity is more likely to have been converted to @placeholder
- Baseline 3: returns the entity that occurred first in the passage. The first appearing entity is arguably the main focus of the passage, hence, more likely to have been also repeated in the question and converted to @placeholder
- Baseline 4: here we begin by extracting all the n-gram (n=2) tokens from the question that contains the token @placeholder. Then, we go through the list of candidate answers, replacing the @placeholder token with each candidate answer for all the extracted n-grams, and counting the number of occurrences of the resulting n-gram. The candidate answer giving the biggest number of occurrences is then returned as the answer.
- SciBERT [25]: a BERT-based language model pre-trained on biomedical scientific corpora. We used the same implementation used by BioMRC.
- CamemBERT [20]: a BERT-based language model pre-trained on French textual corpora. We used the same implementation as SciBERT, as the two models are BERT-based
- FlauBERT [21]: another BERT-based pre-trained language model for French.
- Human performance: we randomly selected 30 instances from the test set, after removing the answers from the instances, we gave them to three non-expert human participants with no biomedical knowledge. They were then instructed to choose an entity from the list of candidate entities as the answer, after reading the passage and the question, even when unsure. The mean accuracy of the three participants is listed in the table of experiment results. Table 3, lists the results obtained with each model, in addition to human accuracy.

The worst performing model is Base 4. The second worst model is Base 1, but, giving the fact that this model randomly selects an entity from the candidate entities as the answer, we can arguably say that this model perform surprisingly well. Going back to Figure 3, we can see that the majority of instances only have

between two and four candidate answers, with the majority having two. This explains the surprising performance of the Base 1 model. In the next update of the dataset, we plan to gather more data, to be able to discard paragraphs with less than four unique biomedical entities. With practically no difference between them and the Base 1 model, came the two French language models, CamemBERT, and FlauBERT. This further confirms what we said about the necessity of having French BQA datasets to be able to advance research and performance in French BQA, as even French monolingual pre-trained language models were not able to surpass the best performing baseline. Base 3 and SciBERT share practically the same score. SciBERT not surpassing the best performing baseline did not came as a surprise to us, this is because it was only trained with English biomedical textual corpora. This fact highlights the need for having a dedicated French biomedical language model trained only or jointly on French biomedical corpora. The best performing model is Base 2, which simply returns the entity most frequently occurring in the passage and the question. Also in the next update of the dataset, after having more data, we plan to discard instances where the most frequent entity in the passage is also the answer. Finally, with 61.11%, non-expert human performance is leading the leaderboard with more than 8% from the Base 2 model. Which suggest that there is ample room for improvement to attain and surpass non-expert human performance. The overall results shows how much work should be done in French BQA, from having more datasets, to having dedicated French biomedical language models.

Table 3: Experiment results

Model	Accuracy
Base 1	44.69
Base 2	52.72
Base 3	46.57
Base 4	41.85
SciBERT	46.86
CamemBERT	44.95
FlauBERT	44.73
Human performance	61.11

5. CONCLUSION

We constructed and made publicly available the first French biomedical question answering dataset, containing 41k+ passage-question instances. The corpus was collected from French biomedical Wikipedia articles, the passages and questions were generated in a cloze-style manner. As a first step towards a leaderboard, we applied and experimented with four baseline models, and three neural-based ones, a biomedical language model, and two French language models. No neural-based model was able to surpass the best performing baseline model, which suggest that in order to face the challenge of French BQA, we need dedicated French biomedical language models trained on French biomedical corpus. With the public release of this dataset along with the leaderboard, we hope to see dedicated French or multi-lingual biomedical language models in the future.

ACKNOWLEDGEMENTS

This work was supported by the Moroccan national center for scientific and technical research (CNRST)

REFERENCES

[1] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2383–2392, doi: 10.18653/v1/D16-1264.

[2] K. M. Hermann *et al.*, "Teaching machines to read and comprehend," in *The 28th International Conference on Neural Information Processing Systems*, 2015, vol. 1, pp. 1693–1701.

[3] E. Choi *et al.*, "QuAC: question answering in context," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2174–2184, doi: 10.18653/v1/D18-1241.

[4] Z. Yang *et al.*, "HotpotQA: a dataset for diverse, explainable multi-hop question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2369–2380, doi: 10.18653/v1/D18-1259.

[5] S. Reddy, D. Chen, and C. D. Manning, "CoQA: a conversational question answering challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019, doi: 10.1162/tacl_a_00266.

[6] M. d'Hoffschmidt, W. Belblidia, Q. Heinrich, T. Brendlé, and M. Vidal, "FQuAD: French question answering dataset," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1193–1208, doi: 10.18653/v1/2020.findings-emnlp.107.

[7] F. A. G. M. T. S. E.-P. S.-M. J. S. R. K. Guillaume Lancrenon Mathilde Bras, "Project PIAF: building a native French question-answering dataset," in *the 12th Conference on Language Resources and Evaluation (LREC)*, 2020, pp. 5481–5490.
[8] Z. Kaddari, Y. Mellah, J. Berrich, T. Bouchentouf, and M. G. Belkasmi, "Biomedical question answering: a survey of methods and datasets," in *2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS)*, 2020, pp. 1–8, doi: 10.1109/ICDS50568.2020.9268742.
[9] H. W. and B. R., "TREC genomics track overview," in *The Twelfth Text Retrieval Conference (TREC)*, 2003, pp. 14–23.
[10] G. Tsatsaronis *et al.*, "An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition," *BMC Bioinformatics*, vol. 16, no. 1, p. 138, 2015, doi: 10.1186/s12859-015-0564-6.
[11] W. L. Taylor, "'Cloze Procedure': a new tool for measuring readability," *Journalism Quarterly*, vol. 30, no. 4, pp. 415–433, Sep. 1953, doi: 10.1177/107769905303000401.
[12] F. Hill, A. Bordes, S. Chopra, and J. Weston, "The Goldilocks principle: reading children's books with explicit memory representations," *In 4th International Conference on Learning Representations (ICLR)*, 2016.
[13] P. H. P. D. Androutsopoulos Ion, "BioRead: a new dataset for biomedical reading comprehension," in *The Eleventh International Conference on Language Resources and Evaluation (LREC)*, 2018, pp. 2771–2776.
[14] S. Kim *et al.*, "A pilot study of biomedical text comprehension using an attention-based deep neural reader: design and experimental analysis," *JMIR Medical Informatics*, vol. 20, no. 1, 2018, doi: 10.2196/medinform.8751.
[15] D. Pappas, P. Stavropoulos, I. Androutsopoulos, and R. McDonald, "BioMRC: a dataset for biomedical machine reading comprehension," in *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, 2020, pp. 140–149, doi: 10.18653/v1/2020.bionlp-1.15.
[16] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program," in *Proceedings AMIA Symposium*, 2001, pp. 17–21.
[17] A. Ben Abacha and D. Demner-Fushman, "A question-entailment approach to question answering," *BMC Bioinformatics*, vol. 20, no. 1, p. 511, 2019, doi: 10.1186/s12859-019-3119-4.
[18] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu, "PubMedQA: a dataset for biomedical research question answering," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 2567–2577, doi: 10.18653/v1/D19-1259.
[19] R. Leaman, R. I. Dogan, and Z. Lu, "DNorm: disease name normalization with pairwise learning to rank," *Bioinformatics*, vol. 29, no. 22, pp. 2909–2917, Aug. 2013, doi: 10.1093/bioinformatics/btt474.
[20] L. Martin *et al.*, "CamemBERT: a tasty French language model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7203–7219, doi: 10.18653/v1/2020.acl-main.645.
[21] V. S. M. C. B. L. A. A. B. C. L. B. Loïc Vial Jibril Frej and D. S. H. Le, "FlauBERT: unsupervised language model pre-training for French," in *The 12th Language Resources and Evaluation Conference*, 2020, pp. 2479–2490.
[22] T. Wolf *et al.*, "Transformers: state-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45, doi: 10.18653/v1/2020.emnlp-demos.6.
[23] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.
[24] A. Tchechmedjiev, A. Abdaoui, V. Emonet, S. Zevio, and C. Jonquet, "SIFR annotator: ontology-based semantic annotation of French biomedical text and clinical notes," *BMC Bioinformatics*, vol. 19, no. 1, p. 405, 2018, doi: 10.1186/s12859-018-2429-2.
[25] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3613–3618, doi: 10.18653/v1/D19-1371.