MedDialog-FR: a French Version of the MedDialog Corpus for Multi-label Classification and Response Generation related to Women's Intimate Health

Xingyu Liu,[1,3] **Vincent Segonne,**[2] **Aidan Mannion,**[1,4]
Didier Schwab,[1] **Lorraine Goeuriot,**[1] **François Portet**[1]
[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
first.last@univ-grenoble-alpes.fr
[2]Université Bretagne Sud, UMR CNRS 6074, IRISA, F-56000 Vannes, France
first.last@univ-ubs.fr
[3]Shesmet [4]EPOS SAS

Abstract

This article presents MedDialog-FR, a large publicly available corpus of French medical conversations for the medical domain. Motivated by the lack of French dialogue corpora for data-driven dialogue systems and the paucity of available information related to women's intimate health, we introduce an annotated corpus of question-and-answer dialogues between a real patient and a real doctor concerning women's intimate health. The corpus is composed of about 20,000 dialogues automatically translated from the English version of MedDialog-EN. The corpus test set is composed of 1,400 dialogues that have been manually post-edited and annotated with 22 categories from the UMLS ontology. We also fine-tuned state-of-the-art reference models to automatically perform multi-label classification and response generation to give an initial performance benchmark and highlight the difficulty of the tasks.

Keywords: Medical Corpus, Women's Intimate Health, Multi-label Question Classification, Response Generation

1. Introduction

Medical conversation data is an essential resource for advancing healthcare research on dialogue systems. However, freely-available medical conversation data in certain languages is often limited, which poses a significant challenge for researchers working in those languages. This is particularly true for certain specialised domains, such as women's intimate health, where the data sources are scarce and data collection is challenging due to ethical and privacy concerns.

There are large-scale dialogue corpora available in the field of healthcare in both English and Chinese, such as *MedDialog* (Zeng et al., 2020) and medical_conversation (Song et al., 2020), two extensive medical dialogue datasets covering various medical specialities. Regarding French corpora, we are aware of only two examples of dialogue datasets. In the context of the *PVDial* project, Campillos-Llanos et al. (2020) created a virtual patient for medical education purposes. Interacting with both clinicians and non-clinicians, they released *PG-logs-eval*, a dataset comprising 115 dialogues. The dialogues simulate medical consultations. While this dataset is clearly useful for studying lexical choices and dialogue, it is based on a virtual agent and not real human patients. More recently, Laleye et al. (2020) introduced a medical conversation corpus of 41 dialogues, *Labforsims*, as part of the development of a dialogue system between virtual patients and physicians. It is also worth mentioning the PxCorpus (Kocabiyikoglu et al., 2023)(Kocabiyikoglu et al., 2022) composed of spoken dialogues between a smartphone and 55 participants including clinicians for drug prescription in French. Although these datasets clearly serve to enrich the French-speaking community, they are too small in size to train data-driven systems. Furthermore, none of them include topics related to women's intimate health, while it is known that in many societies, the latter are disadvantaged by discrimination rooted in sociocultural factors (Aleksanyan and Weinman, 2022; Mehta et al., 2022).

In this study, we address the challenge of limited healthcare dialogue data in the French language by building a corpus of 20,000 dialogues on general medicine and women's intimate health[1]. Due to the lack of available data in French, we leveraged the *MedDialog-EN* English dataset Zeng et al. (2020) and translated it into French. We then post-edited and annotated the translated corpus to support experiments of a multi-label classification task and a response generation task in French [2].

The paper presents (1) the construction of a new medical dialogue dataset in French related to women's intimate health and general medicine

[1]The corpus is available at our Zenodo repository: https://doi.org/10.5281/zenodo.10889881.

[2]The code for the experiments can be found at: https://github.com/getalp/FRMedDialog.

Proceedings of CL4Health 2024 @LREC-COLING-2024, pages 173–183

(in section 2), (2) the implementation and comparison of state-of-the-art approaches for two tasks: multi-label classification and response generation (in section 3).

2. The French MedDialog dataset

In this section, we present the method used in constructing the corpus, including the post-editing of the machine translation of dialogues selected and the manual annotation of questions. Additionally, we provide statistics on the corpus, such as the number of dialogues selected, the analysis of post-editing results, as well as the distribution of topics covered.

2.1. Method

2.1.1. Data Selection and Translation

The source of our corpus, *MedDialog-EN* dataset, contains 257,454 English consultations between patients and doctors. Each consultation is composed of a of textual single-turn dialogue: a patient describing their medical condition and asking a question which is answered by a physician.

MedDialog-FR-women We extracted dialogues concerning women's intimate health through the use of specific keywords provided by 2 women's intimate health practitioners. Each keyword corresponds to a distinct entry in the UMLS Metathesaurus (Bodenreider, 2004), an ontological medical data resource that combines multiple terminology systems.The corresponding keywords and UMLS entries are presented in the appendix Appendix A.

Using a set of 17 keywords, we extracted a total of 16,149 dialogues. However, it should be noted that some of the dialogues extracted were not related to women's intimate health. For instance, the patient identified menopause as her health condition in her question, however, the query pertains to her liver issues, thus falling outside our study's focus.

MedDialog-FR-general In order to provide a more general-domain set of dialogues alongside the focus on women's intimate health, we extracted and translated an additional set of 7,120 dialogues based on a broad set of medical keywords, of which 500 were subsequently post-edited. We consulted with a French physician involved in the field of medical informatics research to put together a list of health conditions judged to be particularly interesting and/or important for general medical practice. The keyword list used to select the general-domain documents from *MedDialog-EN* is detailed in the appendix Appendix A.

The subsequent step involved the use of neural machine translation to automatically translate the chosen dialogues. We employed DeepL's API[3] to automatically translate the extracted dialogues.

2.1.2. Post-Editing

Machine translation has greater productivity and reduced costs compared with human translation. Current state-of-the-art automated translation from English to French appears to yield comprehensible and fluent translations. However, when it comes to medical text, accuracy and clarity are critical.

In order to align the translated text more closely with human standards of accuracy, fluency, and naturalness, and to ensure the accurate translation of medical terminology, we performed post-editing with native French speakers on a portion of the translated text.

PE Platform The post-editing was conducted with *doccano* (Nakayama et al., 2018) (see Figure 1), an open-source web-based text annotation tool. It provides annotation features for text classification, sequence labelling, and sequence-to-sequence tasks. *Doccano* met our requirements in this regard.

Participants In Figure 1, the presented example illustrates a machine translation that translated the English phrase "on birth control" as "*sous contrôle de naissance*" in French, a literal translation that was not natural. A more natural and accurate French expression should be "*sous contraception*". We decided that, for a dataset on women's health, female post-editors would be a more suitable choice. In our case, we recruited 2 undergraduates and 4 master's students in the fields of linguistics or natural language processing with a proficient level (B2) of English, all female. Before the PE task, we provided a one-hour training session to the post-editors with our PE guidelines. At the end of this session, we provided them with five common translations to post-edit, in order to ensure their understanding of the relevant principles.

Guidelines To ensure the effectiveness of the post-editing process and make the target text as consistent as possible, we prepared guidelines for the post-editing task. The basic PE rules in our guidelines rely upon the guidelines established by TAUS (Translation Automation User Society) (TAUS, 2016), such as "Ensure that no information has been accidentally added or omitted" and "Use as much of the raw MT output as possible". Our guidelines also included procedural steps relating to the use of our PE platform and a list of

[3] https://www.deepl.com/api

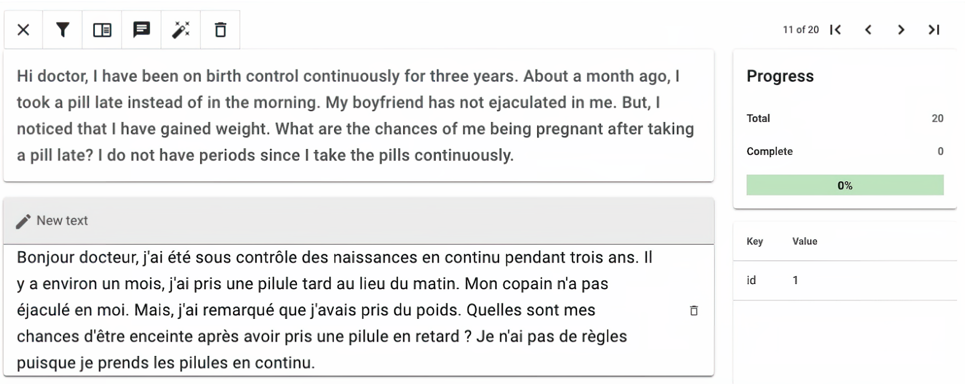

Figure 1: Post-Editing Interface with *Doccano* with an example from *MedDialog-EN*.

	Ed	TER	BLEU
Women's health	37.81	0.079	0.91
General medicine	31.73	0.065	0.92

Table 1: Post-editing effort indicators

corresponding English to French translations for the specialized medical acronyms found in texts. The PE guidelines will be given as supplementary material in the final version of the paper.

PE Technical Effort Indicators Technical effort refers to the alterations made by the translator, which typically include insertions and deletions (Krings, 2001). Once post-editors finished post-editing, following previous works dealing with post-editing (Koponen, 2016; Snover et al., 2006; Alvarez et al., 2020), we calculated the following metrics as PE technical effort indicators (shown in Table 1) : Edit Distance (Ed) calculates the smallest number of edits needed to match the machine translation output with its post-edited version, Translation Edit Rate (TER) quantifies the edit operations required on the word level, and BLEU assesses the coherence of the machine-translated text with the post-edited content. A detailed explanation and examples are provided in the appendix Appendix B.

Lower TER scores indicate better machine translation quality, and higher BLEU scores are generally associated with better machine translation quality, which can lead to reduced post-editing effort. In our task, TER scores below 0.1 and BLEU scores above 0.9 indicate that the results of machine translation were acceptable in general. In terms of the qualitative analysis, the main edits involved modifying medical acronyms and addressing incomplete translations.

Anonymization During the post-editing process, it came to our attention that certain user first names and doctors' names had not been properly anonymized. In order to enhance data privacy, we carried out some additional anonymization steps. Names in questions were replaced with #Person1#, and names in answers with #Person2#. Additionally, URLs, email addresses, telephone numbers and other digits present in the original dataset were identified using regex and replaced by specific strings (e.g. #URL#, #EMAIL#). Manual verification was then carried out on a randomly-sampled subset of question-response pairs.

2.1.3. Annotation of MedDialog-FR-women

We initiated the process of multi-label annotation with questions related to women's intimate health. As for the general medicine data, we plan to conduct the annotation in the future based on labels provided by doctors as per their requirements.

With the goal of categorizing user questions into different themes related to women's intimate health, we leveraged the post-edited data by annotating it with predefined labels provided by 2 women's intimate health practitioners.

Annotation Platform The multi-label annotation task was conducted with *doccano* (See Figure 2).

Participants The question labelling annotators were the same 6 annotators as for PE task. Another training session for the annotation was provided following the PE session.

Labels Following domain-expert recommendations, we selected 27 labels for annotation. This includes 25 topics related to women's intimate health, such as *endometriosis*, *menopause*, and *contraception*, alongside two additional labels: *hors-sujet* (out of scope), indicating questions unrelated to women's intimate health topics, and *autres* (other), signifying women's intimate health subjects not covered by the other labels.

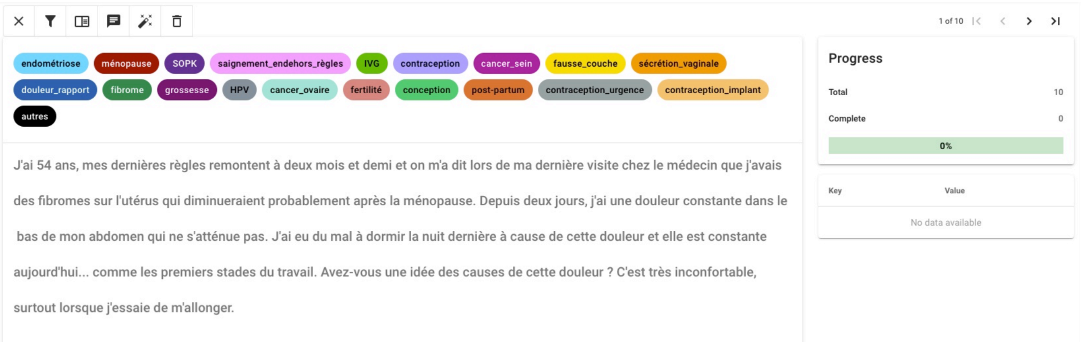

Figure 2: Multi-label Annotation Interface with *doccano*

Guidelines Multi-label annotation refers to annotation shemas where each question can be assigned to multiple labels or categories. In our guidelines, we presented the annotation procedure, provided explanations for each label, and offered illustrative examples to clarify situations involving multiple labels. We also asked annotators to specify the additional categories in the comments when selecting the *autres* (other) label for cases not covered by predefined categories.

2.1.4. Post-Processing of Labels

After instructing annotators to specify the category of the question in comments when selecting the *autres* (other) label, we subsequently processed these comments to consolidate the list of labels and introduce new ones into our dataset.

Additionally, due to the limited occurrence of certain labels, we merged them to create more broadly defined categories, such as combining *contraception_implant* (contraceptive implants), *contraception_urgence* (emergency contraception), and *pillule* (contraceptive pill) into *contraception* (contraception), a single, higher-granularity label.

The post-processing of labels was validated by an expert in women's intimate health. In the end, our multi-label dataset contained 22 labels (showed in Table 2). In the final dataset, we retain both the initial labels and the post-processed labels.

2.1.5. Data Partitioning

We split the *MedDialog-FR-women* multi-label dataset into a training set of 500 instances, a validation set of 100 instances and a test set of 300 instances. The ratio was chosen to balance the need for maximizing the amount of fine-tuning data available while also ensuring that the test set is large enough for the results to be statistically significant, given the scarcity of some categories. To maintain consistent label distribution, we leveraged the iterative stratification algorithm (Sechidis et al., 2011) during the data splitting process. The label distribution is shown in Figure 3.

labels-en	labels-fr
endometriosis	endométriose
menopause	ménopause
PCOS	SOPK
conception	conception
painful sex	douleur_rapport
contraception	contraception
disorders of breast	affection_sein
ovarian cancer	cancer_ovaire
vaginal discharge	sécrétion_vaginale
abortion/VTP	IVG
uterine fibroid	fibrome
fertility/infertility	fertilité
cervical cancer	cancer_col_utérus
abdominal pain	douleur_abdominale
menstruation disorders	menstruation
swelling	gonflement
hot flushes	bouffée_chaleur
emotional disorder	troubles_humeur
out of scope	hors_sujet
sexually transmitted infections	IST
pelvic inflammatory disease	affection_appareil_génital
pregnancy, childbirth or the puerperium	g_a_p

Table 2: 22 post-processed French labels with English translation

2.2. Statistics

To construct our French *MedDialog* Dataset (*MedDialog-FR*), we initially extracted from *MedDialog-EN* and automatically translated a total of 16,149 dialogues related to women's intimate health and an additional 7,120 dialogues related to general medicine. From this dataset, we randomly selected 900 dialogues on women's intimate health and 500 dialogues concerning general medicine for the PE task. Table 3 shows the statistics of the post-edited data. Subsequently, we performed multi-label annotation on the 900 questions extracted from these same dialogues focused on women's intimate health. In total, 1,286 labels were distributed over the 900 dialogues, averaging 1.43 labels per instance.

The 6 annotators were thus tasked with post-editing 900 dialogues related to women's intimate health and 500 dialogues concerning general medicine. Additionally, they annotated in to-

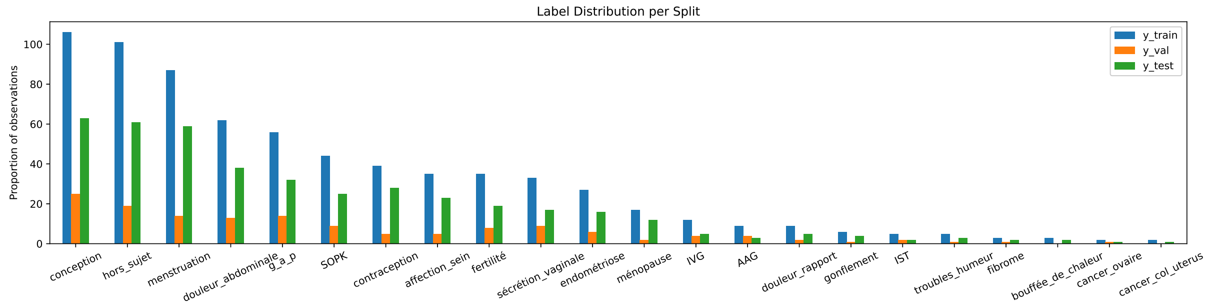

Figure 3: Label distribution per split

	Women	General	Total
# Dialogues	900	500	1400
# Tokens	199,574	81,958	281,532
Avg. # of tokens/Dialogue	225	163	203
Max. # of tokens/Dialogue	897	320	897
Min. # of tokens/Dialogue	59	40	40
Med. # of tokens/Dialogue	162	151	159

Table 3: Size of post-edited data

tal 900 questions from the post-edited dialogues on women's intimate health with multi-labels. The summary of the dataset is shown in Table 4.

Task	Women	General
Machine translation (# dialogs)	16,149	7,120
Post-editing (# dialogs)	900	500
Multi-label annotation (# questions)	900	-

Table 4: Statistics of MedDialog-FR dataset

3. Experiments

In this section, we introduce two tasks carried out using the MedDialog-FR-women dataset: multi-label question classification and response generation. We undertook these tasks with future applications in mind. These processes could be beneficial for determining the topic of a user's question and proposing responses within a dialogue system focused on women's intimate health.

3.1. Multi-label Question Classification

3.1.1. Method

Multi-label classification is concerned with categorizing instances into multiple classes at the same time. Each class associated with a given instance is referred to as a label. Following the previous work (Nam et al., 2014), we adopted the Binary Relevance method to convert the multi-label classification challenge into multiple single-label classification tasks. This approach involves treating each label as an independent binary classification problem.

3.1.2. Models

Our classification architecture comprised a pre-trained BERT type model and a linear layer to convert the BERT representation to a classification task. The `[CLS]` representation is fed into a linear classification layer. We furthermore utilized a binary cross-entropy loss over sigmoid output (`BCELosswithlogits` with *PyTorch*) to measure the error for each label. Given the label imbalance in the dataset, we also experiment with `weighted` versions of the loss function, which aims to balance the precision-recall tradeoff by multiplicatively weighting positively-labelled examples in proportion to their prevalence in the training data.

As for the baseline models, we ran experiments on state-of-the-art large language models for the French language: *FlauBERT* (Le et al., 2020) and *CamemBERT* (Martin et al., 2020); and specialized French models tailored for the biomedical domain: *CamemBERT-bio* (Touchent et al., 2023) and *DrBERT-4G* (Labrak et al., 2023).

3.1.3. Results

We present the results of our experiments involving two labeling approaches: one using all 22 categories and the other restricted to the 12 most common, grouping the 11 least frequently occurring labels under the *autres* (other) category. The purpose of using the 12 labels was to assess the performance of our method on a less imbalanced dataset, where we grouped the 11 least frequently occurring labels under the *autres* (other) category.

22 labels Once the model training was finished, its output could be construed as a probability distribution across the labels for a given instance. Based on its probability in the output, a threshold was required to decide if a label was predicted. We conducted a correlation analysis between various thresholds and F1 scores on the validation set using the best checkpoint during training to establish an optimal threshold for each model, as shown in Figure 4 for example. For the `CamemBERT-bio-base-weighted` model,

Model	macro			weighted
	P	R	F1	F1
`FlauBERT-base`	0.36	0.43	**0.38**	0.59
`FlauBERT-base-weighted`	0.41	0.36	0.37	0.54
`CamemBERT-base`	0.23	0.33	0.26	0.58
`CamemBERT-base-weighted`	0.38	0.29	0.32	0.53
`CamemBERT-bio-base`	0.33	0.40	0.35	0.59
`CamemBERT-bio-base-weighted`	0.45	0.44	<u>0.42</u>	0.63
`DrBERT-4gb`	0.45	0.29	0.33	0.50
`DrBERT-4gb-weighted`	0.40	0.31	0.31	0.46

Table 5: Model performance with on the MedDialog-FR-women test set containing 22 labels

when the threshold was 0.41, we achieved the highest F1 scores on the validation set.

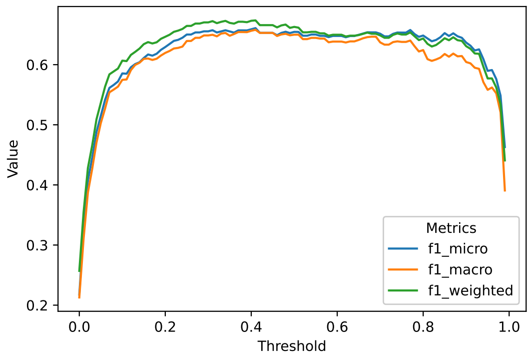

Figure 4: F1 scores of `CamemBERT-bio-base-weighted` on the *MedDialog-FR-women* validation set of 22 labels with different thresholds

Table 5 shows the performance of the models on the annotation test set of 22 labels. Models with the "weighted" suffix in their names indicate their utilization of class weighting in the loss function. The precision, recall and F1 score are calculated as the macro average across all labels. Additionally, the weighted F1 scores are calculated. The two best-performing models are `CamemBERT-bio-base-weighted` and `FlauBERT-base`, with macro-F1 scores of 0.42 and 0.38 respectively.

However, even for these two models, the F1 scores for certain labels with low occurrences, such as *cancer_ovaire* (ovarian cancer) and *troubles_humeur* (mood disorders), are exceedingly low and even reached 0. In an effort to mitigate the impact of data imbalance, we subsequently aggregated the 11 least frequently occurring labels under the *autres* (other) label and evaluated our approach on the remaining 12 labels with `CamemBERT-bio-base-weighted` and `FlauBERT-base`.

12 labels Table 6 shows the performance of `CamemBERT-base-weighted` and `FlauBERT-base` on the test set of 12 labels.

Model	macro			weighted
	P	R	F1	F1
`FlauBERT-base`	0.54	0.60	0.56	0.58
`CamemBERT-bio-base-weighted`	0.56	0.67	0.60	0.61

Table 6: Model performance on the MedDialog-FR-women test set containing 12 labels

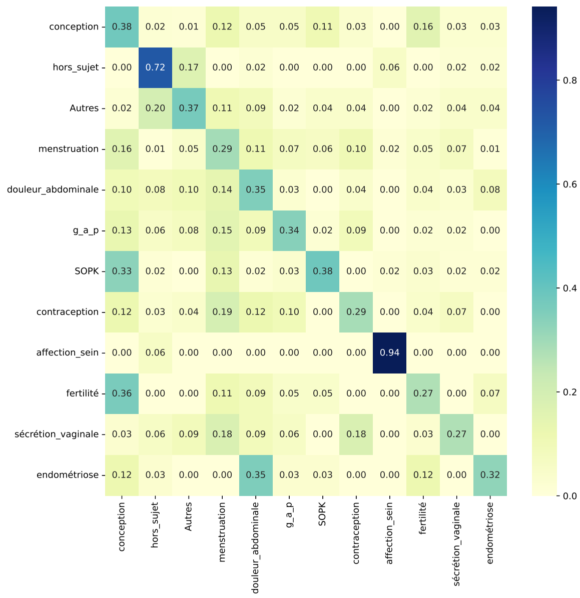

Figure 5: Confusion matrix for CamemBERT-bio-base-weighted on MedDialog-FR-women test set of 12 labels

In contrast to the 22 labels, the macro-averaged metrics for the 12 labels show significant improvement, which is understandable given the reduction in label imbalance. Moreover, the weighted-average scores are similar to macro-average scores, which suggests that the models are not significantly biased towards the larger labels and perform consistently across our dataset.

To assess the model's performance for each label individually, we used the confusion matrix on the predictive accuracy of `CamemBERT-base-weighted` (see Figure 5). The matrix layout consists of rows representing true labels and columns representing predicted labels. High values along the diagonal indicate the model's proficiency in making accurate predictions.

By examining the matrix, we can observe that the *affection_sein* (breast disorders) category has yielded the highest number of accurate predictions. This can be attributed to the relatively independent nature of breast disorders within the domain of women's intimate health, where questions of this category are less related to other topics. Additionally, there is some confusion between *fertilité* (fertility) and *conception* (conception), two closely related concepts. It can be difficult even for humans

Model	PPL ↓	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	Meteor ↑	BertScore ↑
Barthez	3.5	**30.6**%	**18.6** %	**16.1**%	**20.9**%	**70.1**
mBarthez	2.4	27.3%	6.8%	13.2%	18%	68.9
LLama2 (FT)	**1.1**	23.9%	4.5%	13.2%	15.8%	62.0

Table 7: Results of the response generation task using Barthez and LLama2 (w/t Fine-Tuning) on the test set of the corpus.

to clearly distinguish them apart.

3.2. Response Generation

3.2.1. Method

We also evaluated our corpus in a response generation task in which the goal is to automatically generate a response to a given user's question. We followed the experimental protocol described in Zeng et al. (2020) where language models are used to generate answers.

We evaluated several language models from two families. On the one hand, we used the pretrained seq2seq models Barthez (French-only) and mBarthez (multilingual) (Kamal Eddine et al., 2021). We also included LLaMa2[4] (Touvron et al., 2023), an autoregressive model pretrained on a multilingual corpus. For the response generation task, we fine-tuned each of the pre-trained models on the training post-edited and non-post-edited data and evaluated them on the same validation and test post-edited data splits as used in the multi-label question classification task. For inference, we generated responses to questions through a beam search with (n=5) and a top-k random sampling (Fan et al., 2018) set to 50. To measure the quality of our generated answers, we compared them to the gold answers (i.e provided by real doctors) applying standard automatic methods used in generation and automatic translation: perplexity, ROUGE score (Lin, 2004) , METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020). These metrics capture different aspect of the quality of the generated answers: perplexity measures the quality of language modeling, ROUGE and METEOR are used in machine translation to evaluate the similarity between the hypothesis and the reference through n-gram matching. BERTScore measures the similarity between sentences using the BERT language models' representations.

3.2.2. Results

The results of our experiments on the response generation task are presented in Table 7. We also provide examples of generation by different models in Table 8. Overall, the models fine-tuned on our corpus are able to generate well-formed and coherent responses, which is reflected in good perplexity scores (the lower the better). The best model in this regard is Llama2, with an average perplexity of 1.1 on the entire test set, while Barthez obtains the highest perplexity with 3.5.

When examining the evaluation metrics obtained from automatic translation, they consistently behave and reveal two trends: Firstly, the overall results are relatively low, with the best ROUGE-1 score reaching approximately 30%, underscoring the challenging nature of the task with this dataset. Secondly, the Barthez model consistently outperformed other models across all metrics, while LLama2 consistently performed the poorest. Specifically, when comparing the performance of Barthez and LLama2 in terms of BERTScore, Barthez's output appears to be more semantically aligned with human responses than LLama2's. This observation may suggest the challenge faced by general multilingual autoregressive models in adapting to specialized domains and tasks.

4. Conclusion and Future Works

In this research, we present the MedDialog-FR dataset, a French version of the MedDialog-EN dataset. This dataset comprises 16,149 dialogues related to women's intimate health topics and an additional 7,120 dialogues covering general medicine. Within this dataset, 1,400 dialogues have been post-edited, and 900 questions have been annotated with multiple labels pertaining to women's intimate health topics. The post-editing process required several adjustments to align the data more closely with the French medical context, including changes in the names of medications and clinical procedures to their French equivalents. Additionally, the names within the dataset have been converted to French names to better adapt to the target medical system and enhance anonymization. To the best of our knowledge, this dataset represents the first dedicated resource in French for dialogues related to women's intimate health.

Applying current state-of-the-art model to the multi-label classification task exhibits moderate performance with the best model achieving a 60% macro F1 score when dealing with 12 labels. Given the relatively small size of the dataset and the imbalance present in the data, these outcomes can be considered reasonable. They underscore the challenge that current NLP models face when tackling this task.

Automated response generation remains a challenging task, with the highest METEOR score in our experiments being 20%. Overall, the performance of the generative models in these experi-

[4]We used the Llama2-7b-chat version

User question	Bonjour docteur, j'ai beaucoup d'irritation sur l'ouverture de mon vagin et quand je le gratte, ça fait du bien mais ça devient plus irritant. Je suis allée chez le médecin récemment et j'ai eu une infection urinaire, mais je ressens toujours de fortes démangeaisons et je commence à m'inquiéter. J'ai 22 ans et pour l'instant je ne prends que des antibiotiques pour l'infection urinaire.
Human response	Bonjour, cela ressemble à un furoncle. Et c'est surtout lié à l'infection urinaire (pièce jointe supprimée pour protéger l'identité du patient). Continuez à prendre des antibiotiques. Il peut être en phase de guérison et c'est pourquoi les démangeaisons doivent être présentes. Mais vous pouvez prendre du sirop Cital 2 fois par jour pendant 15 jours. Et vous pouvez aussi appliquer de la pommade Candid-B qui aidera à réduire les démangeaisons. Ne vous grattez pas. Cela augmenterait l'infection. Buvez plus d'eau. Gardez la zone vaginale aussi sèche que possible. Si vous avez des pertes blanches par voie vaginale, vous pouvez prendre les comprimés Candid V6 par voie vaginale tous les soirs pendant 6 jours. Si l'infection n'est pas contrôlée par les antibiotiques donnés, il peut être nécessaire de changer d'antibiotiques ou d'utiliser des antibiotiques par injection."
barthez response	Bonjour, d'après vos antécédents, il semblerait que vous ayez développé une candidose, une infection génitale. Faites faire votre culture d'urine pour détecter la bactérie causant l'infection et prendre des antibiotiques en conséquence. Vous pouvez prendre des comprimés contre les infections urinaires jusqu'à deux fois par jour."
mBarthez response	"Bonjour, il pourrait s'agir d'une bactérie. Faites-moi savoir si je peux vous aider davantage. Pour plus d'informations, consultez un obstétricien et un gynécologue en ligne –> https://icliniq.com./ask-a-doctor-online/obstetrician-and-gynaecologist."
Llama2	"Bonjour, je suis désolé d'apprendre que vous ressentez des démangeaisons et de l'irritation à l'ouverture de votre vagin. Il est important de consulter un médecin pour obtenir un diagnostic précis et un traitement adéquat. Il est possible que votre infection urinaire soit la cause de vos démangeaisons et de l'irritation, mais il est également possible que cela soit lié à d'autres facteurs, tels que des infections vaginales ou des allergies. Il est donc important de consulter un médecin pour obtenir un diagnostic précis et un traitement adéquat. En ce qui concerne les antibiotiques que vous prenez pour l'infection urinaire, il est important de suivre les instructions du médecin et de prendre le médicament jusqu'à la fin de la prescription, même si vous vous sentez mieux. Cela aidera à éliminer l'infection et à vous aider à vous sentir mieux.

Table 8: Examples of generated responses to questions from the validation dataset

ments fell short of providing a solid basis for meaningful human evaluation of the generated outputs Nevertheless, a brief examination of some of the generated responses revealed that existing automatic metrics may not be suitable for assessing response generation accurately. As a result, future work will involve implementing more refined fine-tuning techniques to attain outputs that are usable, and subsequently, we intend to conduct thorough human evaluations to more comprehensively assess the models' performance.

To further enable improved model performance, a key component of our upcoming efforts will be to expand our annotations, paying special attention to the less common categories that have low occurrences. We also aim to introduce annotations for medical entities within our datasets, with the aim of making significant contributions to the advancement of task-oriented medical dialogue systems.

5. Acknowledgements

This research was supported by Association nationale de la recherche et de la technologie (ANRT) under grant CIFRE 2021/0693 and was partially supported by MIAI@Grenoble-Alpes (ANR-19-P3IA-0003). We would also like to thank the participants to the annotation process.

6. Ethics Statement and Limitations

Access to actual medical data is heavily restricted in France. We thus used an already publicly available corpus in English. In addition to translation, a number of additional steps were taken to ensure that the MedDialog-FR dataset is fully anonymized and properly adapted for French-language applications. We first made sure that no personal information could be found in the data. This is why we replaced all names that could have been kept in the original data. We also performed post-edition after automatic translation to adapt the phrasing and medical terminology to more natural French. In addition, our annotation project adhered to strict ethical guidelines, which include, but are not limited to, fair compensation for annotators. We do not foresee any direct social consequences or ethical issues.

The primary focus of this study centers on the dataset, with the conducted experiments serving as an initial benchmark to assess the task's complexity. Our preliminary goal is to reach a decent value on automatic metrics such as BLEU and METEOR prior to allocating valuable human resources for output evaluation. These experiments are designed to highlight the tasks' challenges, and we plan to undertake human evaluations once our models attain more robust metrics.

Since the original corpus is derived from dialogues in the U.S.A., there might be some cultural differences with French-speaking countries in the way people interact with doctors and which treatments and medical advises can be provided.

7. Bibliographical References

Yeva Aleksanyan and Jason P. Weinman. 2022. Women, men and covid-19. *Social Science & Medicine*, 294:114698.

Sergi Alvarez, Antoni Oliver, and Toni Badia. 2020. Quantitative analysis of post-editing effort indicators for NMT. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 411–420, Lisboa, Portugal. European Association for Machine Translation.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. PubMed PMID: 14681409; PubMed Central PMCID: PMC308795.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis. 2021. BARThez: a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ali Can Kocabiyikoglu, François Portet, Prudence Gibert, Hervé Blanchon, Jean-Marc Babouchkine, and Gaëtan Gavazzi. 2022. A spoken drug prescription dataset in french for spoken language understanding. In *LREC 2022*.

Maarit Koponen. 2016. Is machine translation post-editing worth the effort? a survey of research into post-editing and effort. *The Journal of Specialised Translation*, 25(2).

Hans P Krings. 2001. *Repairing texts: Empirical investigations of machine translation post-editing processes*, volume 5. Kent State University Press.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. Drbert: A robust pre-trained model in french for biomedical and clinical domains. *bioRxiv*, pages 2023–04.

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 Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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.

Kajal Mehta, Hana Arega, Natalie L. Smith, Kathleen Li, Emma Gause, Joohee Lee, and Barclay Stewart. 2022. Gender-based disparities in burn injuries, care and outcomes: A world health organization (WHO) global burn registry cohort study. *The American Journal of Surgery*, 223(1):157–163.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Jinseok Nam, Jungi Kim, Eneldo Loza Mencía, Iryna Gurevych, and Johannes Fürnkranz. 2014. Large-scale multi-label text classification—revisiting neural networks. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II 14*, pages 437–452. Springer.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*, pages 145–158. Springer.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

TAUS. 2016. Taus post-editing guidelines.

Rian Touchent, Laurent Romary, and Eric de La Clergerie. 2023. CamemBERT-bio: a tasty french language model better for your health. *arXiv preprint arXiv:2306.15550*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

8. Language Resource References

Leonardo Campillos-Llanos, Catherine Thomas, Eric Bilinski, Pierre Zweigenbaum, and Sophie Rosset. 2020. Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation. *Natural Language Engineering*, 26(2):183–220.

Ali Can Kocabiyikoglu, François Portet, Prudence Gibert, Hervé Blanchon, Jean-Marc Babouchkine, and Gaëtan Gavazzi. 2023. Px-Corpus : A Spoken Drug Prescription Dataset in French for Spoken Language Understanding and Dialogue. 10.5281/zenodo.10080490.

Fréjus AA Laleye, Gaël de Chalendar, Antonia Blanié, Antoine Brouquet, and Dan Behnamou. 2020. A french medical conversations corpus annotated for a virtual patient dialogue system. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 574–580.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Appendix A. Matching UMLS Terms for Keywords

Keyword	UMLS Concept
endometriosis	Endometriosis (C0014175)
menstruation/period	Menstruation(C0025344)
menopause	Menopause (C0025320)
PCOS	Polycystic Ovary Syndrome (C0032460)
spotting	Metrorrhagia (C0728993)
conception	Fertilization (C0015914)
pregnancy	Pregnancy (C0032961)
pain&sex/pain&intercourse	Dyspareunia (C1384606)
contraception	Contraceptive methods (C0700589)
breast cancer	Malignant neoplasm of breast (C0006142)
ovarian cancer	Ovarian neoplasm (C0919267)
white discharge/vaginal discharge	Vaginal Discharge (C0227791)
miscarriage	Spontaneous abortion (C0000786)
abortion/VTP	Induced abortion (procedure) (C0392535)
postpartum	Postpartum Period (C0086839)
uterine fibroid	Uterine Fibroids (C0042133)
fertility/infertility	Female infertility (C0015895)
papillomavirus/HPV	Human Papillomavirus (C0021344)

Table 9: Keywords for women's health dialogue selection and corresponding UMLS entries

Keyword(s)	UMLS Concept
angioedema	Angioedema (C0002994)
hypertension, high blood pressure	Hypertension or high blood pressure (C3843080)
hypoglyc(a)emia	Hypoglycemia (C0020615)
ACS, acute coronary syndrome	Acute Coronary Syndrome (C0948089)
pulmonary (o)edema	Pulmonary Edema (C0034063)
cardiac arrythmia	Cardiac Arrythmia (C0003811)
diabet(es|ic)	Diabetes (C0011847)
ketoacidosis	Ketoacidosis (C0220982)
meningitis	Meningitis (C0025289)
cholecystitis	Cholecystitis (C0008325)
pyelonephritis	Pyelonephritis (C0034186)

Table 10: Keywords for general-domain dialogue selection and corresponding UMLS entries

Appendix B. Post-Editing Example

The example below show the difference between a translated sentence and a post-edited one:

Machine translation: *Mes règles ont été retardées de 5 jours. D'habitude, j'ai des cycles réguliers. Je prends de la metformine 1000 mg depuis 6 mois, après avoir appris que j'ai un problème de PCOS. J'ai fait un test de grossesse aujourd'hui matin, mais le résultat était négatif...*

Post-edited: ***J'ai un retard de règles** de 5 jours. D'habitude, j'ai des cycles réguliers. Je prends de la metformine 1000 mg depuis 6 mois, après avoir appris que j'ai un problème de **SOPK**. **Ce matin**, j'ai fait un test de grossesse, mais le résultat était négatif…*

- Ed: Levenshtein distance calculates the minimum number of single-character edits (insertions, deletions or substitutions). For this example, 58 single-character edits are needed to transform the machine translation text into the reference text. This measure provides a granular view of the textual differences, reflecting the extent of similarity or divergence at the character level.
- TER: the Translation Edit Rate (TER) measures human edits on the machine translation, including the insertion, deletion, and substitution of single words. Using the post-edited text as reference, the TER score is calculated as the number of edits needed to change the machine translation into the reference, divided by the total number of words in the reference. There are 13 edits in the example above involving substitution (like "PCOS" (English acronym of Polycystic Ovary Syndrome → "SOPK" (French version)), deletion (such as "aujourd'hui" (today) to say "this morning" in a natural way in French), and insertion (including "J'ai", "un", "retard", to express "I have a late period" in French). The reference text contains 45 tokens, thus, for this single example, the TER score is 13/45 ($\sim$0.289). We also calculate the average TER across all instances.
- BLEU: BLEU measures n-gram correspondence between the machine translated and the reference text. BLEU typically considers n-grams from 1 (unigrams) to 4 (4-grams). For example, the phrase "Je prends de la" (a 4-gram) appears in both texts and will contribute to the 4-gram precision. For each n-gram, the score respresents the number of matching n-grams in the machine translation and reference, divided by the total number of n-grams in the machine translation. In this example, the BLEU score was 0.688.