A Virtual Patient Dialogue System Based on Question-Answering on Clinical Records

Janire Arana†, Mikel Idoyaga†, Maitane Urruela†, Elisa Espina‡
Aitziber Atutxa†, Koldo Gojenola†
†HiTZ Center - Ixa, Bilbao School of Engineering
‡Faculty of Medicine and Nursing, Basurto
University of the Basque Country UPV/EHU
{jarana021, midoyaga002, murruela002}@ikasle.ehu.eus,
{elisa.espina, aitziber.atucha, koldo.gojenola}@ehu.eus

Abstract

In this work we present two datasets for the development of virtual patients and the first evaluation results. We firstly introduce a Spanish corpus of medical dialogue questions annotated with intents, built upon prior research in French. We also propose a second dataset of dialogues using a novel annotation approach that involves doctor questions, patient answers, and corresponding clinical records, organized as triples of the form *(clinical report, question, patient answer)*. This way, the doctor-patient conversation is modeled as a question-answering system that tries to find responses to questions taking a clinical record as input. This approach can help to eliminate the need for manually structured patient records, as commonly used in previous studies, thereby expanding the pool of diverse virtual patients available. Leveraging these annotated corpora, we develop and assess an automatic system designed to answer medical dialogue questions posed by medical students to simulated patients in medical exams. Our approach demonstrates robust generalization, relying solely on medical records to generate new patient cases. The two datasets and the code will be freely available for the research community.

Keywords: virtual patient, question-answering, dialogue understanding

1. Introduction

Virtual patients (VP) have emerged as powerful tools in medical education and health simulation. VPs allow medical students to simulate a real clinical consultation, enabling them to reproduce a wide variety of consultation types, thus gaining valuable experience before medical exams or interviews with real patients. Virtual patients are based on dialogue systems, which are AI-based automated systems designed to exchange information with users through natural language conversations. The main goal of a dialogue system is to enable effective communication between humans and computers, understanding user input in the form of text or voice and to appropriately respond to user demands. Figure 1 presents the main components of a dialogue system:

- The Natural Language Understanding (NLU) module, composed of the following elements:
  - Intent classification (IC): it processes the user's input and predicts their intention (or intent), trying to understand what the user is asking the system to do.
  - Response Location (RL): in charge of extracting the appropriate response from knowledge bases or external documents (in our case, from clinical reports).
- Dialogue Management (DM): responsible of maintaining the consistency of the dialogue context and deciding a response based on the current dialogue state and the user's intent.
- Natural Language Generation (NLG): it creates a response, depending on the result of the response location module.

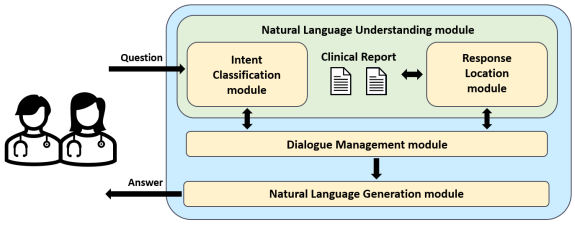

Figure 1: General architecture of a Dialogue System.

In order to train AI to facilitate the development of VPs, well-documented resources and accurate medical dialogues are needed. The aim of this work is to develop the basis for a Virtual Patient in Spanish, focusing on the NLU component. The created system is composed of the Intent Classification and Response Location modules that, having a clinical report and a set of questions as input, is capable of identifying what the user is asking, and then extracting the text fragments where

LREC-COLING 2024, pages 2017–2027

the answers to the questions are located. The applications for the presented datasets are vast, given that these types of data are difficult to access and costly to develop, specially in languages other than English. These are the main contributions of this work:

- We present a corpus of medical dialogue questions in Spanish annotated with intents, derived from the work for French from Laleye et al. (2020).
- We have also annotated a corpus of manually aligned doctor questions, patient answers, and the corresponding clinical record. We present a novel approach to provide the answers given by the VP: instead of having a manually created structured patient record (Campillos-Llanos et al., 2020), our corpus will consist of doctor-patient dialogues aligned in triples of the form *(clinical record, question, answer)*. Thus, an important novelty of our approach lies in the fact that the dialogues are linked to health records that describe a medical episode. This link will allow creating new dialogues from clinical cases, given the difficulty of having access to actual dialogues, while in previous works either a very reduced and limited set of patients has been created by hand, or a set of dialogues has been created for a single patient. The present work shows a way to considerably enlarge the number of possible patients, because thousands of available medical records can be used to simulate virtual patients, widening the applicability of these systems to the training of medical students. Grounding patient dialogues on health records will overcome the bottleneck of having a limited number of patients and will allow to generate lots of different patient profiles based on detailed descriptions found in the medical records.
- Using the annotated corpora, we will develop and evaluate an automatic system to provide the answers of the dialogue questions posed by medical students. We will see how our approach helps to generalize well, needing only a set of medical records to generate new virtual patients.

We must stress that the application to a language other than English has implied a considerable amount of work including translation, manual curation and annotation. Currently, the availability of annotated datasets for medical dialogues in languages other than English is very scarce, and this work makes an important contribution in this respect, opening the way for the training of multilingual medical dialogue systems.

2. Related work

Conversational agents in medicine have long been an object of research (Milne-Ives et al., 2020), as they could support a variety of activities, including behavior change, treatment, health monitoring, triage, and screening. Specifically, virtual patients are a learning tool to prepare students for clinical environments (Isaza et al., 2018), as shown in Figure 2.

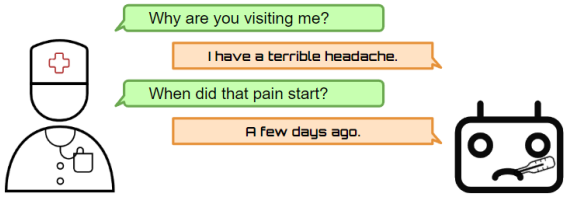

Figure 2: Example of a conversation with a virtual patient

Intent classification is an important part of dialogue systems, which consists in cataloging each question or statement with a category from a predefined set, characterizing its main purpose and finding the intent or goal behind a given utterance. For example, Rojowiec et al. (2020) present a dataset and an evaluation of an automatic intent classifier for clinical questions in German. Laleye et al. (2020) focused on the intent identification task, building a corpus of medical conversations in French. The corpus consists of 41 interviews, made up of 1818 sentences. To achieve the intent-classification, they trained a model based on *FastText* vector representations of words, introducing rule-based constraints. Its main limitation is that all the interviews correspond to variants of just one medical case.

Campillos-Llanos et al. (2020) have designed a dialogue system that handles different specialties and clinical cases. To develop the task they created a patient registration model, a knowledge model and a termino-ontological model with structured thesauri with linguistic, terminological and ontological knowledge. Their approach uses a rule-based methodology and utilizes terminology-rich resources to manage medical interviews. However, this approach, although working well for a reduced set of manually created patient profiles, lacks the generalization that we will intend with the QA approach developed in this work.

Zini et al. (2019) developed a specialized chatbot for OSCE (Objective Structured Clinical Examination) exams using deep learning techniques. A distinguishing feature of this work is that they pose the dialogue between medical students and VPs as a question-answering (QA) task (Singhal et al., 2023) over a natural language text describing the patient's condition, thus greatly simplify-

ing the system design, and also avoiding the time-consuming work of manually creating patient profiles, as any detailed description (e.g. electronic health records) of a patient's status could be used to represent new cases. To train the embedding model they used a corpus of medical documents, obtaining a QA accuracy of 81%. A main drawback of the system is that, for each question, it tries to output a judgment y_i for each sentence s_i, where $y_i = 1$ if s_i is a correct answer for a question and 0 otherwise, that is, the answer corresponds to an entire sentence, that in many cases can add lots of non-relevant context to the specific answer, specially in natural language texts that can contain long sentences or also when the answer to a question is divided in several consecutive sentences.

Chen et al. (2022) present the IMCS-21 corpus, a large-scale medical conversation corpus extracted from the Chinese online health community *Muzhi*, which provides professional medical advisory services for patients. The authors use neural models to perform different tasks, thereby studying the practicality and usefulness of the corpus. The corpus contains annotated information like NER, intent types and diagnoses, although it is not prepared for a VP approach.

Fareez et al. (2022) present a corpus for OSCE exam preparation. This corpus was created by a group of final year medical students in Canada. Medical conversations in English were recorded following the format of the OSCE exams and there were 272 simulated cases between doctors and patients. The audio recordings were transcribed, manually corrected for speech-to-text errors, and speaker identifiers (D for physician and P for patient) were added. The resource most relevant to the work presented are the dialogues, which can be used to train a NLP/QA model to replace traditional standardized patients for OSCE with a virtual patient.

In the last years, there has been a big leap in machine reading comprehension, which has become a central task in natural language understanding, with large-scale datasets (Hewlett et al., 2016; Joshi et al., 2017) and a diverse set of QA architectures (Wang et al., 2017; Huang et al., 2018). Recent work has produced systems that surpass human-level accuracy on the Stanford Question Answering Dataset (SQuAD), one of the most widely-used reading comprehension benchmarks (Rajpurkar et al., 2016).

It can be concluded that there are diverse approaches to the task of building conversational systems, with different techniques and combinations of modules. It should be noted that in this work an integration of a reading comprehension QA system and an intent classification model have been chosen, but these modules are not necessarily present in all approaches. For example, Zini et al. (2019) focus exclusively on the QA module, with a sentence-based approach, while Laleye et al. (2020) approach the intent classification task. Other works (Chen et al., 2022; Fareez et al., 2022) are entirely concerned with the development of datasets and corpora to help train and evaluate downstream models.

3. Resources and Methods

In this section we will first present the development of the two annotated corpora we have created in subsection 3.1, and then we will describe the experimental design of an automatic intent classifier and the QA-based virtual patient in subsection 3.2. The corpora and the code and parameters of the experiments will be freely available[1].

3.1. Corpora

In this work, two main tasks will be carried out: intent classification and question answering. To achieve this, it has been necessary to develop two different corpora, one for each task:

- VIR-PAT-INTENTS: a corpus composed by 2691 doctor utterances annotated with their corresponding intent category, used to train an intent-classification model.
- VIR-PAT-QA: a corpus composed of 129 doctor-patient consultation dialogues and the clinical reports corresponding to such consultations, amounting to a total of 6290 question-answer pairs, used to train an extractive question answering model.

3.1.1. The VIR-PAT-INTENTS corpus

VIR-PAT-INTENTS is an adaptation from the French corpus of Laleye et al. (2020). The corpus was automatically translated from French to Spanish and it was manually corrected, giving 2691 utterances where 145 different intent types were identified. They were classified in a hierarchical way from more generic to more specific. Initially we took the intent set from the French corpus as inspiration, but we found that the intent classification was made upon a single patient case with multiple dialogues. This implies that many questions were related specifically to the (unique) current illness, not taking into account the possibility of dealing with questions corresponding to different symptoms or diseases. For that reason, we extended this set to better generalize and give a more detailed account in order to obtain a wider and more general intent classification. We also made changes into the original hierarchy in order

[1] https://github.com/Midoiaga/VirPat-2024

to ease the task of generating more natural answers. We think that this will be specially useful in the response generation phase, as the intent type can add additional information about the subject or the type of answer, which can be helpful to generate more natural responses. For example, the intents *Symptom_patient* and *Symptom_family* should be answered differently, as the first one asks for a response in first person, the patient himself, while the second type would need third person plural because it refers to the patient's family. Similarly, we distinguished *Answer_YesNo* and *Answer_Describe*, which ask for a short or detailed answer, respectively.

Table 1 shows the 11 main categories, which are subdivided in more specific groups, giving a four level hierarchy in the most specific subcategories (see Table 2). Defining a four level hierarchy does not mean that all the main categories branch until the last level. Some categories, such as *greeting*, *goodbye*, *others*, *affirm* and *state* do not have more subcategories, because the utterances in the corpus do not have a more specific intention other than the main one. For the rest of the categories, the subdivisions follow a specialization depending on each higher level category. Levels 3 and level 4 are used to specify the type of question posed by the doctor, extending the theme of the previous branch. For example, in the second level of *symptom*, the subcategories are related with the individual that is suffering the asked symptom (*family, environment, patient*). Then, the following branches of *patient* are related to different aspects of the patient's current disease, such as symptom localization, start, pregnancy... For example, the doctor could ask different aspects of the patient's disease: "Where is the pain located?" (*localization*), "How often do you have the symptoms?" (*frequency*), "When did the symptoms start?" (*start*), etc.

The *treatment* category contains the deepest trees of the hierarchy, where the most specific subcategories are related to different types of treatment a patient can receive. It is worth mentioning that the deepest branches of most of the categories end with *describe* and *yes_or_no*, because adding those subcategories allows to specify more natural conversations. For example, *yes_or_no* is used when the speaker expects a yes/no answer to the question and *describe* when they are waiting for a longer explanation.

The corpus is divided in three different sets in a stratified way, 80% for training containing 2079 utterances and 10% for development and test with 306 utterances each.

3.1.2. The VIR-PAT-QA corpus

The corpus presented in Fareez et al. (2022) was used as the basis for the development of the VIR-PAT-QA corpus. It is composed of doctor-patient consultation dialogues in English, recorded following the format of the OSCE exams and then transcribed and manually corrected. Our aim is to associate each dialogue with a corresponding clinical record in natural language, enriching the dialogue dataset with a textual description of each patient, thus opening the way to create new virtual patients simply by adding new clinical records to the dataset. The creation of the corpus was performed following different steps (see Figure 3):

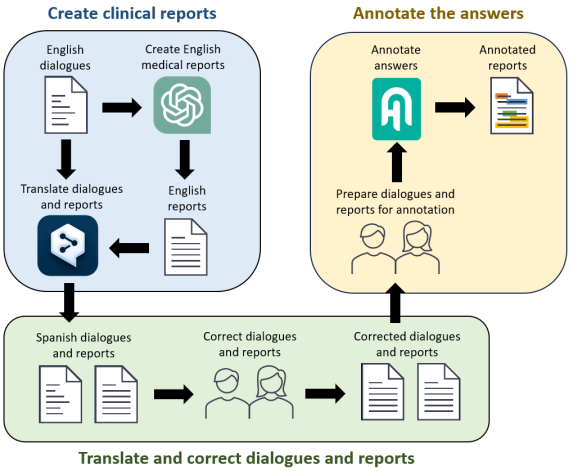

Figure 3: Creation of the QA corpus.

- Create clinical records. Given the transcribed dialogues, a generative model[2] was used to create the corresponding reports, taking as input the dialogue together with a prompt explaining the main sections of a clinical record.
- Translate and correct dialogues and reports. Both dialogues and reports were translated into Spanish and manually corrected to avoid translation errors. Although the quality was initially good, there were some problems which were corrected, including:
  - Errors in number, gender or mixing formal and informal addressing. We adapted expressions, politeness issues and we corrected certain gender biases. For example, in English the term *patient* refers to both female and male patients while in Spanish the article or pronoun is different depending on whether the patient is male or female, and in many cases the automatic translation engine did not keep track of the dialogue flow translating a part of the dialogue as if the

[2] https://chat.openai.com/

Main Categories	Description	Examples	Amount
afirmar (affirm)	Afirmative utterances.	*Sí* (Yes)	13
despedida (goodbye)	Parting utterances.	*Adiós* (Bye)	6
estado (state)	General condition questions	*¿Cómo estás?* (How are you?)	7
motivo_de_consulta (reason for consultation)	Reason of the visit	*¿Qué te trae por aquí?* (What brings you here?)	61
otros (others)	Utterances that do not belong to other categories	*Le pido su tarjeta sanitaria* (I ask for your health card)	5
personal (personal)	Questions about patient's life	*¿Cómo te llamas?* (What is your name?))	499
psiquiatria (psychiatry)	Questions about patient's feelings	*¿Eres feliz?* (Are you happy?)	60
saludo (greeting)	Greeting utterances.	*Hola* (Hi)	25
sintoma (symptom)	Questions about patient's symptoms	*¿alergias?* (Alergies?)	1604
tratamiento (treatment)	Questions about patient's received treatments	*¿otros medicamentos?* (other medications?)	384
vida_sexual (sexual life)	Questions about patient's sexual life	*¿Es usted sexualmente activo?* (Are you sexually active?)	27
Total			**2691**

Table 1: Description of the main intent categories.

Intents per hierarchy level		
Level	**#categories**	**Examples**
1	11	affirm symptom personal psychiatry
2	34	personal_sports personal_sleep psychiatry_mood symptom_family symptom_patient
3	111	psychiatry_mood_describe psychiatry_mood_yes/no symptom_patient_fever treatment_operation_results
4	145	personal_addiction_alcohol_frequency personal_addiction_smoke_amount symptom_patient_localization_describe treatment_consultation_specialist_yes/no

Table 2: Number of intents per hierarchy level.

patient was a female and the rest as if the patient was a male. In some cases the profession of the patient could lead the automatic translator to assume a gender that was not actually explicit in the original health record. A similar problem happens with politeness formulas. In Spanish there is a polite way of communicating using the "usted" form. The automatic translator did not take into account the fact that old people tend to use the polite form while young people do not.

- Missing information, when the information in some answers given by the patient was not present in the clinical record.

Overall, the quality of the corpus has been a main concern. We have been exhaustive while producing the annotated corpora, and both dialogues and reports were manually corrected and double-checked to avoid translation errors, also devoting an effort to keep the language of the dialogues as natural as possible.

- Annotate the answers. Given the question-answer pairs in dialogues and the corresponding clinical record, each question was linked with the matching answer text in the clinical record. The questions were classified into two main categories, one of them with two subcategories:
  - *Questions that need to be answered*: questions that require seeking the answer in the reports.
    - *Answered questions*: the response appears in the report.
    - *Unanswered questions*: when the information necessary to respond does not appear in the report, the span is empty and the *is_impossible* attribute value was set to *True*. This type corresponds to questions where the patient did not understand the question or when the answer in the dialogue did not answer the question. In fact, the participants in the dialogue generation process in (Fareez et al., 2022) were instructed to respond similarly to patients in a clinical/hospital setting, with vague responses to open-ended questions and specific responses to direct questions.
  - *Questions that do not need and answer*, as when the medical student makes a comment. For example, in a relatively important proportion of the utterances given by the doctor or medical student there are expressions like "Thank you", "OK", "Now I will check your temperature" that do not require an answer from the patient. It is important that these types of sentences are understood and detected, because otherwise a standard QA system would try to give an answer for every question it is being asked, and that would be incorrect for these types of utterances.

After this process, the questions appearing in the dialogue are linked to the span of text in the clinical record that answers the question. Figure 4 shows an example of an annotated dialogue, where some answers span over a single word while others can contain longer explanations.

The final dataset contains a total of 6290 question-answer pairs from 129 different clinical cases in the SQuAD v2.0 format Rajpurkar et al. (2016). It was divided into training, development and test

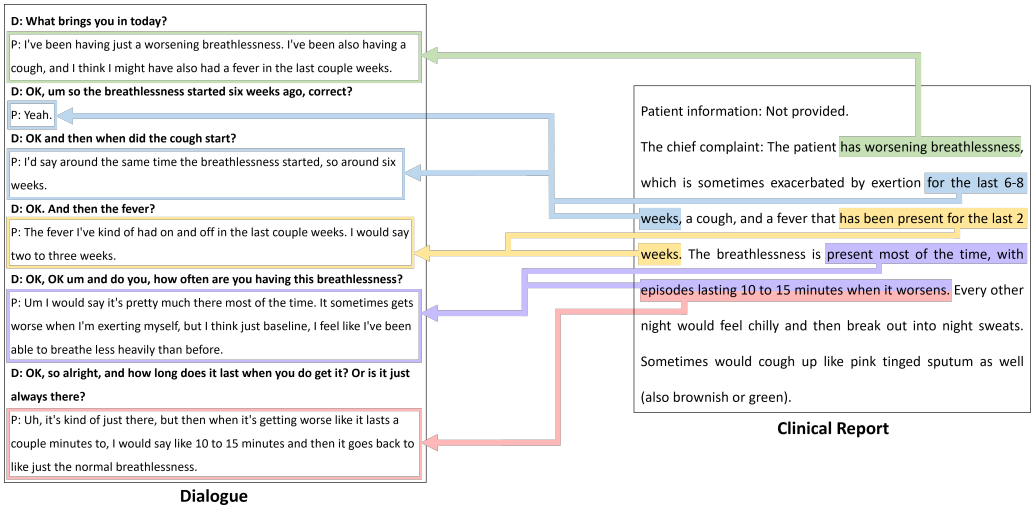

Figure 4: Example of a dialogue annotated together with its corresponding clinical record. Each question and answer are linked with the text in the clinical record containing the answer.

sets (75%, 10% and 15% of the corpus respectively). Table 3 shows the number of questions of each type per set. The table shows that an important percentage of questions do not contain an answer in the given clinical record, an important aspect presented in Rajpurkar et al. (2018), as extractive QA systems can tend to make unreliable guesses on questions for which the correct answer is not stated in the context.

After several annotation rounds and different refinements of the annotation guidelines, the annotators reached a good level of agreement, with an exact match of 65.98, and a total agreement (partial + exact) of 88.94, where we considered a partial match when the information content in both annotations was the same, even if they differed in a single token or a non-essential modifier.

Question type	train	dev	test
Questions that need to be answered	4573	496	915
Answered questions	2753	295	580
Unanswered questions	1820	201	335
Questions that do not have to be answered	228	27	51
Total	**4801**	**523**	**966**

Table 3: Number of questions by type in train, dev and test sets

3.2. Experimental design

After developing the two datasets for intent classification and QA, our aim was to train a neural-based system for each task (subsections 3.2.1 and 3.2.2, respectively), and also test the effect of the sequential application of the two modules on the extractive QA task (see subsection 3.2.3).

For both tasks, to train a deep learning model from scratch with good results, very large amounts of data and resources would be needed but, although the size of the two generated corpora is far from trivial, it is still insufficient to obtain an accurate model. Since the advent of pre-trained models, ways have been created to refine them more efficiently to perform new tasks. In this case, intermediate tasks or STILTs (Supplementary Training on Intermediate Labeled-data Tasks) have been used (Phang et al., 2018). According to several studies, as shown in Vu et al. (2020), training the model to perform intermediate tasks before performing the target ones can be very beneficial, especially when there is limited data (see Figure 5). The hyperparameters used in the experiments are detailed in Table 10.

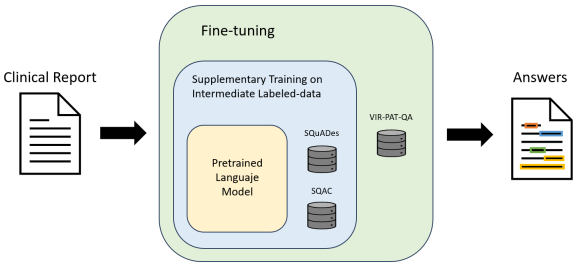

Figure 5: STILT training process.

3.2.1. Intent classification

In a first set of experiments, we will train an intent classifier based on transformers, that is, the objective is creating a model capable of knowing the intention of a user for each utterance. This model is specialized in predicting the intention of doctors' utterances in the context of a virtual patient. Different pre-trained models have been tested trying to find the most suitable one for our goals and we finally chose two models. The first one, *BERTIN*, was trained with general Spanish texts (de la Rosa et al., 2022) based on the BERT architecture, and

the other one with Spanish bio-medical texts called *"bio-bsc-es"* (Carrino et al., 2022) based on the *RoBERTa* architecture. After several initial tests, we saw that the best results were obtained using the full set of labels, that is, performing a classification task over the full set of 145 intent types.

3.2.2. Extractive QA

In this work, the process of getting the answers from the text has been posed as an extractive Question Answering task. For this, once the corpus was collected and labeled, a model based on transformers was trained. Our task consists in learning an extractive QA system over triples of the form (*clinical record*, *question*$_i$, *answer*$_i$) to allow the virtual patient to predict the correct judgment over QA pairs about the clinical record.

As with intent classification, the ideal would be to use a QA model trained on the medical domain but, after an exhaustive search, no such model could be found. However, two models were selected that might be useful for the task: SQuADes[3](Stanford Question Answering Dataset in Spanish) and SQAC[4] (Spanish Question Answering Corpus). SQuADes and SQAC contain 100,000 and 18,800 *(context, question, answer)* triples, respectively. Both models are fine-tuned versions of RoBERTa (MarIA-RoBERTa Gutiérrez-Fandiño et al. 2021) in Spanish, developed by BSC (Barcelona Supercomputing Center). Although they do not correspond to the medical domain, their size is much larger than our QA corpus, so our hypothesis is that they could help to extend the coverage of our system, and using them as intermediate tasks can provide significant benefits in the results, since they share the type of task (extractive QA) and the language (Spanish).

3.2.3. Combination

For a final set of experiments, all the questions from the VIR-PAT-QA corpus were taken and their intention was predicted using the previously developed intent classifier. We performed two different experiments. In the first one, instead of the questions, the input to the QA system was just the intent type, with the aim of testing how much the intent types are adjusted to give a precise description of the question's objective. In the second one, the predicted intents were added to the question, having the intent first, followed by the question text. This will test whether the sequential application of the two systems could help improve the results.

[3] https://huggingface.co/IIC/roberta-base-spanish-squades

[4] https://huggingface.co/IIC/roberta-base-spanish-sqac

3.3. Evaluation methods

For the evaluation of the intent classification system, we have used recall, precision and F1-score. Given the high class unbalance we decided to calculate the macro and weighted average. Macro average (see equation 1[5]) gives equal importance to all the classes while the weighted average calculates the average over all instances, independent of each class (see equation 2).

$$MacroAvg.\ (M) = \frac{\sum_{i=1}^{\#class} M_i}{\#class} \quad (1)$$

$$W.\ Avg.\ (M) = \sum_{i=1}^{\#class} \frac{\#instances_i}{\#instances} M_i \quad (2)$$

To evaluate the performance of the QA-based module, two metrics have been used. On the one hand, the exact match metric (EM), which measures the percentage of answers that have been 100% correct, i.e., the answers that exactly match the gold standard. This is a strict measure which can heavily penalize answers that differ in a single character or token.

We will also make use of a more relaxed metric, the F1-score, a common metric for classification tasks, and also widely used in QA. It is computed over the individual words in the prediction against those in the gold answer. The number of shared words between the prediction and the truth is the basis of this score: precision (see equation 3) represents the percentage of correct words contained in the model response, and recall (see equation 4) is the ratio of the number of shared words to the total number of words in the gold response. The F1-score is a harmonic mean between precision and recall over the set of words contained in each answer (see equation 5).

$$precision = \frac{\#shared\_tokens}{\#predicted\_tokens} \quad (3)$$

$$recall = \frac{\#shared\_tokens}{\#gold\_tokens} \quad (4)$$

$$F1{-}score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (5)$$

4. Results

In the following three subsections we will present the results of the intent classifier, the extractive QA-based system and their combinations.

4.1. Intent classification

In Table 4 we can see the results given by our intent classifier on the test set. Comparing the results of the models, the *bio-bsc-es* model, spe-

[5] M: precision, recall, and f1-score

cialized in medical texts, performed better predicting the intents of utterances that contain medical terms. Several experiments were performed with different preprocessing steps, but the simplest ones, lower case and removing punctuation, obtained better results. The other preprocessing experiments, like removing "stopwords" or stemming, did not get any improvement.

		Precision	Recall	F1-Score
BERTIN	Macro Avg	0.81	0.84	0.81
	W. Avg	0.79	0.80	0.78
bio-bsc-es	Macro Avg	0.88	0.89	0.87
	W. Avg	0.84	0.86	0.84

Table 4: Results for the intent classification module using the *BERTIN* and *bio-bsc-es* models.

4.2. Question-Answering

Table 5 presents the results of the extractive QA-based system using the SQuADes and SQAC models. The table shows the exact score, which measures the proportion of exactly answered questions, and the F1-score, that averages the number of correct tokens per answer, giving a better account of partially answered questions. We have differentiated the overall score on all questions and the results for the questions that contain an answer in the given document. The last row measures the proportion of correctly detected unanswerable questions, that do not contain an answer taking the patient's clinical record as input.

		SQuADes	SQAC
Exact	All	64.70	64.29
	HasAnswer	52.93	51.98
F1	All	73.04	73.94
	HasAnswer	65.70	66.76
NoAnswer		86.87	87.46

Table 5: Results of extractive QA obtained with the fine-tuned SQuADes and SQAC models.

4.3. Combinations

We had interest in also evaluating the contribution of intent types in the QA task. Table 6 presents the results of the QA module taking the intent type as input (upper part of the table), and also the combination of question and intent (lower part). As could be expected, indicating only the intent type the performance is considerably lower than with the natural language question, although it obtains an exact score of 23.93 with SQAC for answerable questions, and a much higher result (82.69) for the non-answerable ones. This can be understood taking into account that, most of the times, knowing just the intent type can be a relevant clue to distinguish questions that need no answer. Including the intent type and the question gives a slight improvement over the question alone.

			SQuADes	SQAC
Intent only	Exact	All	44.20	44.31
		HasAnswer	22.50	23.93
	F1	All	50.87	50.79
		HasAnswer	32.70	33.86
	NoAnswer		85.07	82.69
Intent + question	Exact	All	65.11	65.42
		HasAnswer	53.25	53.57
	F1	All	73.47	74.09
		HasAnswer	66.05	66.83
	NoAnswer		87.46	87.76

Table 6: Results of extractive QA with SQuADes and SQAC using only the intent (upper part of the table) and intent+question (lower part) as input.

5. Discussion

Regarding intent classification, the results in Table 4 show that the models trained on the *bio-bsc-es* corpus give the best results. Examining the details, we have seen that most of the errors come from confusions appearing at the lowest levels of the intent hierarchy. From the 40 errors committed by the automatic system, only 10 of them correspond to errors at the first level, while most errors (26) occurred at the third level of the hierarchy, with 4 errors for levels 2 and 4. Table 7 presents some examples of errors at each level. In the first level, the system confuses a *goodbye* intent with a greeting, or when a single question is asking two different responses (second row). In the rest of the levels (2 to 4), the examples in the table show that the errors occur at the finest level of distinction.

Table 8 presents several errors committed in the QA task (answers marked as partial or incorrect). The first three examples in the table show how most of the times there are slight differences in the answer span, which in many cases do not constitute a problem because the predicted answer partially extends the correct answer. Regarding the incorrect answers, the last two lines exemplify some of the most frequent errors. In the first incorrect example, the question made by the doctor is ambiguous, not precising the exact meaning of the question (… *what about the numbness in the groin area?*). Similarly, in the last example, the predicted answer could also be considered correct, given that the patient refers several problems. To sum up, we see that the QA system is robust enough even in the cases marked as incorrect.

The use of STILTS as an implementation strategy has as one of its advantages the generalization given by a large corpus of *(context, question, answer)* triples, such as SQuADes, with 100,000 instances, compared to our medical dialogue QA dataset (6,290 instances). We tested the generalization ability of the system to new specialties leaving aside the QA pairs of a specialty from the training and development sets, and evaluating on that specialty. We selected the *musculoskeletal* spe-

Level (#errors)	Gold intent	Predicted intent	Examples
1 (10)	*goodbye*	*greeting*	Goodbye sir my best wishes to your wife
	personal_environment_children_yes/no	*symptom_patient_pregnancy_amount*	Do you have children and how many pregnancies have you had?
2 (4)	*psychiatry_mood*	*psychiatry*	You feel confused You are happy by nature
3 (26)	*personal_diet_frequency*	*personal_diet_yes/no*	Do you often eat fast food?
	symptom_patient_history_yes/no	*symptom_patient_yes/no*	have you had stones? Have you had a head injury?
4 (4)	*personal_addiction_smoke_amount*	*personal_addiction_smoke_yes/no*	How much do you smoke?
	symptom_patient_appearance_yes/no	*symptom_patient_appearance_describe*	Does red appear at the time of menstruation?
	symptom_patient_change_yes/no	*symptom_patient_change_describe*	When you put yourself in a position like the fetal position that you curl up on yourself, does that calm the pain a little?

Table 7: Errors in intent classification.

Level (#errors)	Question	Gold answer	Predicted answer
Partial	Did you say that in the morning the stiffness lasts more than 30 minutes?	lasts more than 30 minutes	Morning stiffness lasts more than 30 minutes
	But you don't notice changes anywhere else?	There is no change in the skin in any other area	No
	Any changes in your vision or hearing?	There have been no changes in the senses	Has not had changes in senses, breathing difficulties,
Incorrect	And what about the numbness in the groin area?	Started two months ago	Has numbness in the groin area
	Good morning, how can I help you?	unbearable pain in the hip	presented with a story of falling 2 hours ago down the stairs on her hip

Table 8: Errors in Question Answering (SQuADes).

		Test (all specialties)	Test (musculoskeletal)
Exact	All	64.70	53.58
	HasAnswer	52.93	52.94
F1	All	73.04	68.40
	HasAnswer	65.70	69.07
NoAnswer		86.87	60.78

Table 9: QA results with SQuADes evaluated on a specialty not present in the training set.

cialty with 25 clinical cases and 1,258 QA pairs for the final test. Table 9 presents the results, showing that the system is robust, even when trained on different specialties, maintaining its performance on the answerable questions with the *Exact* measure and a 4.5 point decrease in the F1-score. The main decrease in performance (26 points) comes from the set of unanswerable questions. We hypothesize that this may come from the bigger size of the SQuADes dataset that contains mostly answerable questions, and plan to improve this problem using other resources such as SQuADRUn (Rajpurkar et al., 2018), containing over 50,000 unanswerable questions written adversarially.

6. Conclusions

We have presented a new dataset for the development of virtual patients in Spanish. It is composed of two corpora, one of questions in a dialogue annotated with their intent types and a second one that links each dialogue to its corresponding clinical record and allows to cast the virtual patient as a question-answering task, given that each question and answer are annotated with their corresponding text in the clinical record. This approach can simplify the creation of new patient profiles taking new clinical records as input. As a main result, we have generated a good quality corpus of doctor-patient dialogues based on different clinical cases. The availability of this type of information is very scarce in the medical domain, and this problem is more acute for languages other than the bigger languages like English and Chinese.

We have evaluated the first version of the datasets using them to train an automatic system with promising results. Additional experiments have demonstrated that the system generalizes well to specialties that were not present in the training set. This work has shown that this is a viable approach that will help to extend the array of possible patients, enriching the applicability of medical virtual patients.

It must be noted that the present work is centered on extractive question-answering, where the answer given by the system is the text as it appears in the clinical record answering the question. However, in order to have a natural interaction with the user, the response should be given in first person, instead of presenting the verbatim text in the clinical record. For future work, we plan to develop a natural language generation module trained and evaluated on the set of pairs given by the patient answer in the dialogue and the clinical record text, which are available in the annotated dataset.

Training Hyperparameters			
	Intent classif.	QA SQuADes	QA SQAC
Learning Rate	5e-5	25e-6	5e-5
Weight Decay	0	0.01	0.1
Train Epochs	50	20	10
Lang id	-	5	5
Max Answer Length	-	512	512
Warmup Steps	0.05	10	10
Per GPU Train Batch Size	32	16	16
Gradient Acc. Steps	0	4	4
Max. Seq. Length	40	384	384

Table 10: Parameter and hyperparameter details for intent classification and QA. The other parameters are set to default values from Huggingface 4.20 version.

7. Acknowledgements

This work was partially funded by the Spanish Ministry of Science and Innovation (MCI/AEI/FEDER, UE, DOTT-HEALTH/PAT-MED PID2019-106942RB-C31 and EDHER-MED PID2022-136522OB-C22), the Basque Government (IXA IT1570-22), MCIN/AEI/ 10.13039/501100011033 and European Union NextGeneration EU/PRTR (DeepR3, TED2021-130295B-C31), Euskampus Fundazioa (EUSK22/19), and the EU ERA-Net CHIST-ERA and the Spanish Research Agency (ANTIDOTE PCI2020-120717-2).

8. Ethics Statement

All the models and data used in this project are public. It also respects privacy policies, because all of the clinical reports were created maintaining the anonymity. We are not aware of any negative consequences that can be generated by this work.

9. Bibliographical References

Leonardo Campillos-Llanos, Catherine Thomas, Éric 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.

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pretrained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland. Association for Computational Linguistics.

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2022. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics*, 39(1). Btac817.

Javier de la Rosa, Eduardo G. Ponferrada, Manu Romero, Paulo Villegas, Pablo González de Prado Salas, and María Grandury. 2022. BERTIN: efficient pre-training of a spanish language model using perplexity sampling. *Proces. del Leng. Natural*, 68:13–23.

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, Thomas Lo, and Christopher W. Smith. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9(1):313.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodríguez Penagos, and Marta Villegas. 2021. Spanish language models. *CoRR*, abs/2107.07253.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. WikiReading: A novel large-scale language understanding task over Wikipedia. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1545, Berlin, Germany. Association for Computational Linguistics.

Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. 2018. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. In *International Conference on Learning Representations (ICLR)*.

Andres Isaza, Maria Teresa, Gary Cifuentes Alvarez, and Arturo Arguello. 2018. The virtual patient as a learning tool: A mixed quantitative qualitative study. *BMC Medical Education*, 18.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Fréjus A. A. 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, Marseille, France. European Language Resources Association.

Madison Milne-Ives, Caroline de Cock, Ernest Lim, Melissa Harper Shehadeh, Nick de Pennington, Guy Mole, Eduardo Normando, and Edward Meinert. 2020. The effectiveness of artificial intelligence conversational agents in health care: Systematic review. *J Med Internet Res*, 22(10):e20346.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks. *ArXiv*, abs/1811.01088.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Robin Rojowiec, Benjamin Roth, and Maximilian C Fink. 2020. Intent recognition in doctor-patient interviews. In *International Conference on Language Resources and Evaluation*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Mahdavi, Jason Wei, Hyung Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature*, 620:1–9.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926, Online. Association for Computational Linguistics.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198, Vancouver, Canada. Association for Computational Linguistics.

Julia El Zini, Yara Rizk, Mariette Awad, and Jumana Antoun. 2019. Towards a deep learning question-answering specialized chatbot for objective structured clinical examinations. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9.