# Learning Health-Bots from Training Data that was Automatically Created using Paraphrase Detection and Expert Knowledge

**Anna Liednikova**[&,#]**, Philippe Jolivet**[&]**, Alexandre Durand-Salmon**[&]**, Claire Gardent**[†]

[&] ALIAE

[#] Université de Lorraine

[†] CNRS

{philippe.jolivet,alexandre.durand-salmon}@aliae.io
{anna.liednikova,claire.gardent}@loria.fr

## Abstract

A key bottleneck for developing dialog models is the lack of adequate training data. Due to privacy issues, dialog data is even scarcer in the health domain. We propose a novel method for creating dialog corpora which we apply to create doctor-patient interaction data. We use this data to learn both a generation and a hybrid classification/retrieval model and find that the generation model consistently outperforms the hybrid model. We show that our data creation method has several advantages. Not only does it allow for the semi-automatic creation of large quantities of training data. It also provides a natural way of guiding learning and a novel method for assessing the quality of human-machine interactions.

## 1 Introduction

Current data-driven dialog models require large quantities of training data. Because of privacy issues, the situation is even worse in the health domain where data is particulary scarce. In this work, we propose a novel method for automatically creating the training data necessary to learn a chatbot which can mimick a doctor in doctor-patient interactions. Specifically, we combine expert knowledge provided by physicians with automatic paraphrase extraction techniques. We first ask experts (physicians) to specify typical doctor-patient interactions occurring in the context of clinical studies when talking about the four main topics generally discussed in these studies namely, sleep, mood, anxiety, leisure. Formally, the specification takes the form of a decision tree whose nodes are labelled with either an example doctor question or an example patient input. Each node in the tree is associated with a unique identifier which can be viewed as a simple form of dialog state.

We then enrich this initial dialog data by extracting paraphrases for patient turns from an online forum.

This data generation method has several advantages. First, it allows for a straightforward integration of expert knowledge in data generation, model learning and model evaluation as we can use the dialog turn identifiers both to guide learning and to assess the model (by comparing the sequences it follows with the expert defined sequences). More generally, the association of each dialog turn with a dialog turn identifier which reflects is position in the decision tree and the consistent use of this identifier during data creation, model learning and model evaluation allows for increased interpretability. Second, this method helps achieve good coverage as we can ensure that the data does contain all possible dialogue paths. This is not the case with WoZ and crowdsourcing data collections approaches where the coverage of the possible dialog paths depends on the crowdworker decisions and input. Third, by instantiating each dialogue with different paraphrases, we can increase linguistic diversity i.e., we can create dialogues that have the same structure but different wording.

In sum, our work makes the following contributions. We propose a novel method for creating training data for dialog models. We apply this method to create training data for a bot mimicking doctor-patient interaction in the context of clinical studies. We use the created data to learn a generation and a hybrid classification/retrieval dialog model, we show that the generation model generally outperforms the classification model and we provide a detailed analysis of the models results using automatic metrics, human evaluation and qualitative analysis.

## 2   Related Work

Various methods have been proposed to facilitate the creation of training data for dialog. Previous work has explored Wizard-of-Oz (WoZ) experiments in which two humans interact based on some pre-defined scenario and the dialogs resultings from these interactions are collected (Green and Wei-Haas, 1985; El Asri et al., 2017) or crowdsourcing settings where workers provide continuations to incomplete dialogs (Wen et al., 2017). Both these approaches are time intensive. Crowdsourcing is also expensive while the human-human dialogs that are collected by both these approaches may be very different from the human-machine interactions that should be learned to support efficient human-machine communication where typically, chat messages are restricted in length. Other work has relied on already available dialog data or on question/answer pairs extracted from online forums (Wei et al., 2018; Lin et al., 2019; Xu et al., 2019). In the health domain however, such data is extremely scarce and difficult to obtain. When obtainable, it also requires extensive pre-processing due to annonymisation constraints. Another line of research has been to acquire data through machine-machine simulations (Xu et al., 2019; Majumdar et al., 2019; Shah et al., 2018). In particular, (Majumdar et al., 2019) combines pre-defined dialog outlines with template-based verbalisations of dialog turns to automatically create a synthetic dialog corpus. Our work is similar to (Majumdar et al., 2019) but differs from it in two main ways. First, instead of using templates, we use automatically extracted paraphrases to enrich the initial dialogs. Second we experiment with two dialog models to investigate how domain knowledge (in the form of decision tree positional information) can best be exploited to guide learning and to support error analysis.

## 3   Creating Dialog Corpora

To create training data for the dialog bot, we start by collecting typical dialog outlines from an expert. We then extract paraphrases for the patient turns from a Health forum and filter out dialogue interactions with low coherence.

**Collecting an Initial Corpus from an Expert.**   Studies have shown that the closed questionnaires traditionally used in the context of clinical studies are ineffective in gathering correct and precise information about the patient status. Our long-term goal is to develop a Human-Machine dialog system that would complement standard clinical questionnaires by engaging the patient in a dialog about the questionaire topics. To create our dialog corpus, we therefore asked a physician to formalise typical patient-doctor interactions occurring in the context of a clinical study in the form of a decision tree describing which questions need to be asked and for each question, which answers are possible. The interactions covers four domains namely, sleep, mood, anxiety and leisure activities and the decision tree has 58 nodes. A fragment of the decision tree created for the sleep domain is shown in Figure 1 on the left and an example dialog for the SLEEP domain in the same figure on the right. We call the data collected from the expert $D_{init}$.
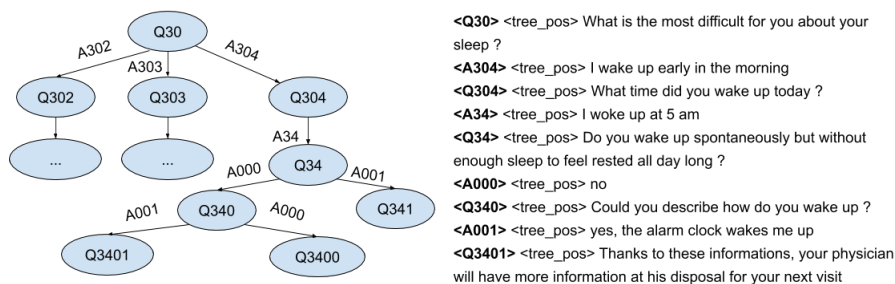


Figure 1: **Fragment of decision tree for the sleep domain and a corresponding dialog**

**Extracting paraphrases.**   We extract paraphrases for the patient turns provided by the expert from the HealthBoard [1] forum in several steps as follows.
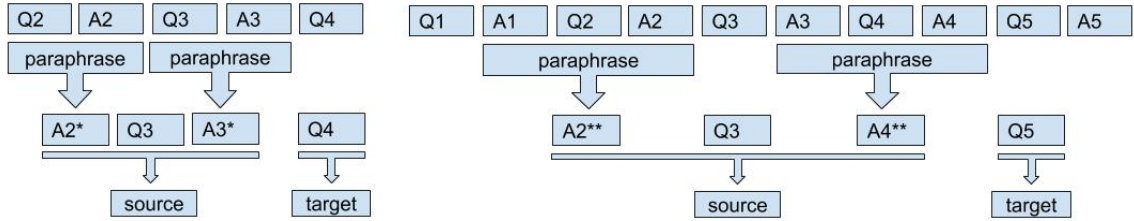
---
[1] healthboards.com

Figure 2: Paraphrasing source-target pairs from $INIT_{long}$ (left) and $INIT_{short}$ (right)

As patient turns are mostly assertive responses to the doctor questions, we start by filtering out questions from the forum data to keep only those utterances which are assertions[2]. To this end, we use a binary stacked Bi-LSTM classifier trained on the Switchboard dataset.

We then compare each patient turn in $D_{init}$ together with its context ($P$, the preceding doctor turn) with the assertive utterances extracted from the forum. For each sequence $D + P$ of contextualised patient turns in $D_{init}$ and each (assertive) utterance $U$ in the forum, we create an S-BERT embedding (cf. Figure 2 left). We then retrieve from the forum all utterances $U$ whose cosine similarity with a contextualized patient turn $D + P$ is higher than 0.70. Finally, we use Maximal Marginal Relevance (MMR) to select from this pool of candidates a subset of paraphrases which maximise both similarity (the paraphrases should be semantically similar to the input turn) and diversity (the resulting set of paraphrases should be maximally diverse[3]. We stop selecting sentences as soon as MMR score becomes negative as a negative MMR score indicates that adding more paraphrases will not increase diversity.

As illustrated in Figure 2, we apply this paraphrase extraction process not only to create paraphrases for a single turn but also to create paraphrases which summarise 3 consecutive turns. In this way, we can derive compressed versions of the initial dialogues. For instance, we can derive the short dialog in (2) from the longer dialog interaction shown in (1).

(1) D1: Do you sleep well ?
    P1: No
    D2: What keeps you awake ?
    P2: I have pain in the legs

(2) D1: Do you sleep well ?
    P1D2P2: No, I have pain in the legs and that keeps me awake.

We refer to the set of paraphrases that summarise three consecutive turns as SHORT and those that summarise a single turn as LONG.

**Filtering Paraphrases.** We compute cosine and BertScore on the S-BERT embeddings of each pair $\langle C, D \rangle$ of context-doctor interactions (where the context is the string concatenation of the three preceding turns) created in the previous step and keep only those pairs for which both scores are higher than the corresponding scores for the corresponding turn in the initial corpus (INIT).

**Training Data.** Table 1 summarises the training data we created. INIT is the dialog data collected from the expert; FORUM , the dataset obtained by replacing patient turns in INIT with their paraphrases, and ALL, the dataset left after filtering. ALL is the combination of LONG and SHORT. As the table shows, the filtered dataset ALL is 5.5 times smaller but has similar coherence (identical or near identical

---

[2]As noted by a reviewer, this is a simplification as in fact, users tend to formulate clarification and disambiguation questions. We leave this for future work.

[3]MMR is a measure for quantifying the extent to which a new item is both dissimilar to those already selected and similar to the target (here the patient turn). It is defined as: $Arg \max_{P_i \in C_U \setminus S} (\lambda \, Sim1(P_i, U) - (1 - \lambda) \max_{P_j \in S} Sim2(P_i, P_j))$ where $U$ is a contextualized user turn, $C_U$ is a pool of candidate paraphrases for $U$, $P_i; P_j$ are paraphrases in $C_U$, and $S$ is the set of already selected paraphrases. A high $\lambda$ value favors similarity. Conversely a low $\lambda$ value results in higher diversity. We set this parameter to be 0.5. We use BertScore recall as function $Sim1$ as this permits checking similarity on a word basis and cosine as function $Sim2$ since we do not need precise comparison between forum sentences, we just want them to be diverse.

|  | INIT | FORUM | ALL | LONG | SHORT |
|---|---|---|---|---|---|
| Nb of src-tgt pairs | 388 | 4 010 696 | 733 104 | 373 220 | 359 884 |
| Nb of distinct turns | 483 | 60 346 | 28 734 | 9 761 | 19 027 |
| Nb of tokens | 18 180 | 204 309 290 | 37 705 130 | 19 546 597 | 18 158 533 |
| Avg Nb of tokens per turn | 13.81 | 12.726 | 12.858 | 13.09 | 12.61 |
| Vocabulary size | 426 | 13169 | 11314 | 10 631 | 9593 |
| Cosine |  | 0.51 | 0.52 | 0.59 | 0.45 |
| BertScore |  | 0.83 | 0.83 | 0.84 | 0.83 |

Table 1: Corpus statistics (INIT: dialog data collected from the expert; FORUM: extension of INIT with paraphrases; LONG: filtered FORUM dataset with only the single turn paraphrases; SHORT: filtered FORUM dataset with only the three-turn paraphrases; ALL = SHORT+LONG)

cosine and BertScore scores) while retaining 86% of the vocabulary and 48% of the unique turns present in FORUM. To facilitate learning and reduce training time, we therefore use the filtered datasets in our experiments.

## 4 Health Bot Models

We aim to learn a model which mimicks a physician in the kind of doctor-patient interaction that is typical of clinical studies conversations.

As we derive the training data from the decision tree, each patient turn and each doctor query is associated with a dialog state (a node in that decision tree). We use this dual information (dialog turn and dialog state) to train and compare two models for response generation: a classification model which, given the last three turns of a doctor-patient interaction, predicts a dialog state and outputs the corresponding doctor query; and a generative sequence-to-sequence model which auto-regressively generates an answer while conditioning on the last three dialog turns. For both models, we use a pre-training and fine-tuning approach similar to that presented in (Radford and Salimans, 2018).

**Classification model.** Given a dialog context (3 dialog turns), the classification model predicts a dialog state and outputs the corresponding doctor query. Thus, the classification model is a multi-class classifier with 58 target classes, the 58 dialog states defined by the expert decision tree. We use the PyTorch implementation of (Radford and Salimans, 2018)'s pre-training and fine-tuning approach provided by Huggingface[4] and the default hyper-parameter settings.

The input to the model consists of three turns $\langle p_1 d_1 p_2 \rangle$. We concatenate these three turns, prefixing each turn with its dialog state identifier and separating them with a delimiter token. Each token is represented by the sum of three embeddings: a word and a position embeddings which are learnt in the pre-training phase; and a turn embedding (learned during fine tuning) indicating whether the token belongs to a patient or to a doctor turn. The input to the model is the sum of all three types – word, position and turn embeddings for each token in the input sequence.

The pre-trained model is the Generative Pre-trained Transformer-based (GPT-2) Language Model trained on the BooksCorpus dataset (7,000 books from different genres including Adventure, Fantasy, and Romance). The parameters are initialized to the smallest version of the GPT-2 model weights open-sourced by (Radford and Salimans, 2018).

We fine-tune the pretrained language model on our data by passing the input turns through the pre-trained model and feeding the final transformer block's activation to an added linear output layer followed by a softmax to predict a probability distribution over the target classes.

**Generative model.** To generate (rather than retrieve) doctor queries, we use the TransferTransfo[5] model (Wolf et al., 2019) which combines a pretrained language model with a Transformer-based generation model fine tuned on dialog data using multi-task learning. Multi-tasking combines a language modeling loss with a next turn classification loss. For the latter, the model is trained to distinguish a correct continuation from one randomly chosen distractor. As for the classification model, we use the

---

[4]https://github.com/huggingface/pytorch-openai-transformer-lm

[5]https://github.com/huggingface/transfer-learning-conv-ai

GPT-2 language model pretrained on the BooksCorpus. For fine-tuning, we use the same augmented representations as for classification i.e., each input consists of the three previous turns with a separator and a dialog state identifier between each turn. From this sequence of input tokens, a sequence of input embeddings for the Transformer is constructed by summing the word and positional embeddings learned during the pre-training phase and the turn embeddings learned during fine-tuning. Multi-task learning is done, as in the TransferTransfo model, by jointly optimizing the language modeling and the next-turn classification loss.

## 5 Experiments

### 5.1 Data and Experimental Setting

We train our models on LONG, SHORT and ALL (cf. Table 1) using a 80/20 train/validation ratio.

We created test data for both long and short interactions by manually specifying six distinct paraphrases for each user turn ($TEST_{LONG}$) or 3 turn sequences ($TEST_{SHORT}$) in $INIT$. Models trained on the ALL dataset are evaluated on $TEST_{ALL}$ which is a concatenation of $TEST_{LONG}$ and $TEST_{SHORT}$. $TEST_{LONG}$ has 4248 source-target pairs and $TEST_{SHORT}$ 2172.

Both models are 12-layer decoder-only transformer with masked self-attention heads (768 dimensional states and 12 attention heads) a dropout probability of 0.1 on all layers (residual, embedding, and attention). They use learned positional embeddings with supported sequence lengths up to 512 tokens. The input sentences are pre-processed and tokenized using bytepair encoding (BPE) vocabulary with 40,000 merges (**?**). Relu activation function is used. CLASSIF is a transformer with a language modelling and a classification head on top, the two heads are two linear layers. The classification head has dropout of 0.1 The model was fine-tuned with a batch size of 8, using OpenAI Adam with a learning rate of 6.25e-5 and a linear learning rate decay schedule with warmup over 0.2% of training. $\lambda$ was set to 0.5. The GEN model is a transformer with a language modelling and a multiple-choice classification head on top, the two heads are two linear layers. The model was fine-tuned with a batch size of 4, using AdamW with a learning rate of 6.25e-5, $\beta 1 = 0.9$, $\beta 2 = 0.999$, L2 weight decay of 0,01. The learning rate was linearly decayed to zero over the course of the training. For both models we trained for 3 epochs using cross entropy loss.

### 5.2 Evaluation

We assess the output of our models using both automatic metrics and human evaluation.

**Automatic Metrics.** In our data, each dialog turn is associated with a node (or dialog state) in the initial decision tree drawn by the expert. We use this dual information (dialog turn and dialog state) for the evaluation. We compute F1 on dialog state labels to analyse the coherence of the system response with the current dialog context (For the generative model, if no label was predicted, the score is 0). We also compute BLEU-4 and BertScore between the model output and the reference turn to assess the similarity of the generated output with the reference.

**Human evaluation.** We ask annotators, stemming from the ALIAE start-up working on health bots and from academia, to interact with a bot which at each new user turn outputs the doctor query suggested by one of our two models. The interaction stops when the bot repeats a previously output question or when the annotator outputs a closing turn (Bye!).

To assess the quality of the bots response given the dialog context, annotators are required to score each system response on a 5 point likert scale with respect to coherence (Is the bot question coherent with the dialog so far ?) where 1 is totally incoherent and 5 is perfectly coherent. For the generation bot, we additionally ask the annotators to rate fluency (Is the bot response well-formed ?) where 1 is unreadable and 5 is perfectly readable. The annotators are non native but their English is fluent. For each model (CLASSIF and GEN trained on LONG), we collect 50 dialogs from 20 annotators. Each annotator interacts at most 5 times with the bot.

We also evaluate the quality of the full dialogs resulting from these human-bot interaction. At the end of each human-bot conversation, the annotator is asked to rate satisfaction on a scale from 1 to 5. In

addition, we applied the evaluation protocole proposed by (Li et al., 2019). Using the 50 dialog pairs collected for bot response evaluation, we show the annotators pairs of collected dialogs, one dialog from the generation model and the other from the classification model and ask them the questions recommended by the protocole: Who would you prefer to talk to for a long conversation? If you had to say one of the speakers is interesting and one is boring, who would you say is more interesting? Which speaker sounds more human? Which speaker has more coherent responses in the conversation? For this task, we had 16 annotators annotating 50 dialog pairs. Each pair was rated 3 times except 2 pairs which were only rated twice. Each annotator annotated at most 10 dialog pairs.

We report the percentage of time one model was chosen over the other. We also compute the average user turn length (number of tokens), the average dialog length (number of turns) and the proportion of turn sequences of length at least two which occur in the decision tree (Sequence Rate). In measuring how often the bot reproduces a sequence of dialog states that is present in the expert decision tree, this latter metrics provides an estimate both of a task success (i.e., how much of the required information has been collected, what proportion of the decision tree has been covered) and how much the collected dialog deviate from the decision tree (how many turns are not about the medical topics covered by the decision tree).

## 6 Results

We compare the classification and the generation models using both automatic metrics and human evaluation. We present various ablation settings to analyse the impact of dialog state information on performance. And we display an example dialog between a human and the generative model in Figure 7.

### 6.1 Automatic Metrics

| Model | F1 | | | BLEU-4 | | | BERTScore | | |
|---|---|---|---|---|---|---|---|---|---|
| | L | S | A | L | S | A | L | S | A |
| CLASSIF Oracle | 0.79 | 0.43 | 0.78 | 0.83 | 0.42 | 0.75 | 0.97 | 0.91 | 0.97 |
| CLASSIF | 0.63 | 0.38 | 0.48 | 0.65 | 0.39 | 0.48 | 0.95 | 0.91 | 0.92 |
| CLASSIF (predict only) | 0.63 | 0.37 | 0.40 | 0.66 | 0.37 | 0.42 | 0.95 | 0.91 | 0.91 |
| GEN Oracle | 0.83 | 0.68 | 0.85 | 0.62 | 0.52 | 0.62 | 0.96 | 0.95 | 0.97 |
| GEN | 0.66 | 0.39 | 0.50 | 0.49 | 0.34 | 0.37 | 0.95 | 0.93 | 0.93 |
| GEN (predict only) | 0.61 | 0.38 | 0.47 | 0.46 | 0.13 | 0.35 | 0.95 | 0.92 | 0.93 |
| GEN (no d-state) | - | - | - | 0.52 | 0.36 | 0.40 | 0.87 | 0.85 | 0.80 |

Table 2: **Results** on **L**ong, **S**mall and **A**ll Datasets

Table 2 shows results for different versions of the generative and classification models depending on which dialog state information is provided in the source and the target, at test and at training time.

In the oracle setting (Oracle), dialog state information is provided for all dialog turns in the input, at training and at test time. This gives an upper bound on how the system would perform given perfect dialog state information. We compare this oracle setting with a standard setting (CLASSIF and GEN) in which only the dialog state associated with the doctor queries are given. At training time, this is the reference dialog state associated with the doctor query. At test time, it is the dialog state of the doctor query predicted by the model.

To analyse the impact of dialog state information on performance, we also execute an ablation study considering models where (i) no dialog state information is given in the input but the model is trained to predict the output dialog state (predict only) and (ii) a model where dialog state information is not used at all (no dialog states).

**Generation outperforms classification.** The F1-score is consistently better for the generation models across all datasets which suggests that learning to generate the system response also helps predicting the correct system dialog state. As regards similarity with the reference, the generation models also consistently show better BERT score but lower BLEU-4 scores. This is coherent with the specificities of each model. Because the generative model generates the system response rather than select it from

the training data (as is the case for the classification model), the similarity in terms of word overlap (as measured by BLEU) with the reference is lower. Nonetheless the high BERT score indicates a strong semantic similarity between the reference and the generated output.

**Predicting the output dialog states helps.** For both the classification and the generation model, dialog state information helps improve performance. As expected the improvment is strongest for the oracle setting. The ablation study further demonstrates that predicting and using predicted dialog state information (CLASSIF, GEN) yields better results compared to settings where dialog state information is only predicted (CLASSIF/GEN predict only) or not used at all (GEN no d-state).

**Shorter interactions are hard to learn.** Contrasting the results from Short and Long in Table 2, we see that scores for the SHORT dataset are lower across the board – it is harder to handle short interactions. This is because, in that setting, the model needs to handle patient turns which convey multiple information – often from different domains – and, based on this, must decide on the correct response i.e. move to the correct dialog state. For instance, in Example (1), the model must (i) detect that the patient turn conveys information about both the sleep and the pain domain and (ii) decide to skip the dialog state corresponding to D2 in example (2).

**Domain analysis.** Table 3 shows the results per domain for the generation and the classification models trained on LONG[6]. Unsuprisingly, results are better for domains (Leisure and Anxiety) with a small number of classes (fewer transitions to learn) and when the training data is larger (Anxiety vs. Leisure and Sleep vs. Mood). This suggests two directions for further research: other paraphrasing techniques could be used to create more training data for those domains where the training data is small and the decision tree drawn by the expert could be refined to yield more balanced domain subtrees.

| Domain | # D-States | % Tg Data | CLASSIF | | | GEN | | |
|---|---|---|---|---|---|---|---|---|
| | | | F1 | BLEU-4 | BERTScore | F1 | BLEU-4 | BERTScore |
| Mood | 18 | 10.44 | 0.36 | 0.55 | 0.90 | 0.64 | 0.54 | 0.95 |
| Sleep | 33 | 63.40 | 0.47 | 0.64 | 0.92 | 0.60 | 0.51 | 0.95 |
| Leisure | 5 | 4.67 | 0.55 | 0.68 | 0.93 | 0.53 | 0.47 | 0.93 |
| Anxiety | 7 | 21.49 | 0.89 | 1.00 | 0.98 | 1.00 | 0.69 | 0.97 |

Table 3: **Results per Domain** (GEN and CLASSIF models trained on LONG)

## 6.2 Error Analysis

We use the expert decision tree to analyse how far off the models predictions are from the correct predictions and compute the proportion of cases where the predicted dialog state is the expected one (Correct), the child of this state in the decision tree (Child Node) or its parent (Parent Node). We also compute the proportion of cases where the predicted and the expected dialog state have the same grand parent (Same Gd Parent) and for all remaining cases whether they occur as different leaves of the tree (Diff. Leaves) and are or are not in the same domain (In Domain, Out of Domain). Table 4 shows the results.

**Most predictions are correct or almost correct.** We find that together the cases where the prediction is almost correct (Child or Parent Node) covers 13.93% and 13.57% of the cases for the generative and the classification model respectively. This means that the prediction of the dialog state is correct or almost correct 76.52% and 80.01% of the time for the classification and the generative model respectively.

**Most errors are an artefact of the decision tree.** Most predictions which are very far off the expected dialog state are transitions associated with the end of the dialog (Diff. Leaves). This is because although turns concluding a dialog are similar for all domains and all dialog paths, they are associated in the dialog tree with different dialog state identifiers. This could be fixed by assigning each leaf node the same identifier and restarting the chatbot using a turn from another domain when reaching such a node. More generally, this shows that alternative design choices for representing the expert knowledge might impact performance.

---

[6]For the other datasets (SHORT and ALL), we observe the same trends.

Interestingly, the use of dialog states derived from the expert dialog tree increases interpretability and allows for a detailed analysis of the errors made by the models suggesting possible directions for improvement such as for instance, using the same dialog state identifier for the end of dialog transitions in all domains and all dialog paths (to reduce the proportion of Diff. Leaves error) and focusing on identifying these factors which would help better differentiate between turns associated with closely related dialog states Child or Parent Node).

| Error Type | CLASSIF | GEN |
|---|---|---|
| Correct | 62.59 | 65.86 |
| Child Node | 4.4 | 3.28 |
| Parent Node | 9.53 | 10.87 |
| Same Gd Parent | 1.69 | 1.31 |
| Diff. Leaves | 15.96 | 11.51 |
| In Domain | 4.28 | 6.65 |
| Out of Domain | 1.53 | 0.26 |

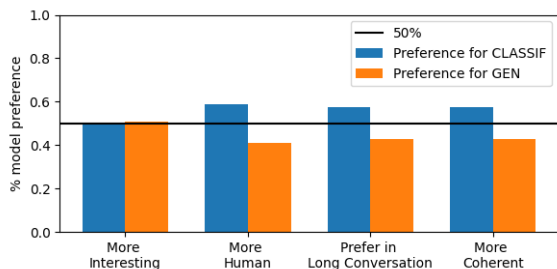Table 4: **Error Analysis on Predicted Dialog States** (GEN and CLASSIF models trained on LONG)



Table 5: **Human Comparison of the two Models**

| Criteria | CLASSIF | GEN |
|---|---|---|
| *Responses* | | |
| Coherence | $3.14 \pm 1.43$ | $3.32 \pm 1.40$ |
| Fluency | - | $4.08 \pm 1.28$ |
| *Whole Dialogs* | | |
| Satisfaction | $3.12 \pm 1.19$ | $3.38 \pm 1.03$ |
| User turn length | $4.92 \pm 3.93$ | $5.04 \pm 4.24$ |
| Dialogue length | $24.42 \pm 3.79$ | $28.98 \pm 5.63$ |
| Sequence rate | $0.35 \pm 0.13$ | $0.26 \pm 0.12$ |

Table 6: **Human Evaluation Scores** (differences are not significant)

| | |
|---|---|
| **Model:** | I am a bot and I will ask you some questions about your health. What is the most difficult for you about your sleep ? |
| **Human:** | I wake up multiple times during the night |
| **Model:** | is there a specific event or a change in your activity which could explain this pain? |
| **Human:** | I believe it's because of my incoming exams |
| **Model:** | are they frequent or infrequent? |
| **Human:** | Only twice a year |
| **Model:** | even if it is not always easy to treat chronic sleep issues why don't you speak with |
| **Human:** | I don't think it's that urgent for now |
| **Model:** | okay! do you have any other things to tell me? |
| **Human:** | I really feel tired all day long |
| **Model:** | would you agree to discuss anxiety and stress? |
| **Human:** | Sure! |
| **Model:** | if you already feel those troubles for the last days why don't you speak with your physician |
| **Human:** | I don't think I really need it for now |
| **Model:** | Do you feel like events have been able to affect your mood lately? |
| **Human:** | My exams |
| **Model:** | do you ever find yourself thinking that it would be best not to wake up the next mornin |
| **Human:** | No, it's just my exams!! |
| **Model:** | do you sometimes think about ending your life? |
| **Human:** | I just told you I'm ok |
| **Model:** | in the long run remaining with such unpleasant thoughts doesn't really seem to me to be ten |
| **Human:** | I think I should be good after my exams are finished |
| **Model:** | Were you rather anxious? |
| **Human:** | Yes, I still am, I don't want to fail |
| **Model:** | What activities you enjoy participating in? |
| **Human:** | I like to play baseball with my university's team |
| **Model:** | Okay! Thank you for sharing. So far I have nothing new to tell you. |

Table 7: **Conversation between Human and Generative Model**.

## 6.3 Human Evaluation and Qualitative Analysis

Tables 6 and 5 show the results of the human evaluation.

**Response quality.** We find that the generative model (GEN, fluency: 4.08) succeeds in generating well-formed responses. Responses that are rated low are often incomplete (e.g., 'in the long run remaining with such unpleasant thoughts doesn't really seem to me to be ten' instead of 'tenable'). This is likely due to the model learning an average sentence length which is below that of longer turns and could be remedied by improved tuning. Both models provides reasonably coherent answers (Coherence: 3.14 and

3.32) and while the generative slighty edges out the classification model, the difference (we used a t-test) is not statistically significant at $p < 0.05$.

**Dialog quality.** Dialogs are quite long which indicates that the bot succeeds in driving a non trivial conversation with the user.

We also observe that the user turns are much shorter than in our training dataset because annotators often respond to questions by a simple yes or no statement rather than a full sentence. This raises the question of how to encourage the user to be more collaborative and provide more informative responses. We leave this here as an open question for further research.

The sequence rate is around a third for each model. Recall that this metrics is the ratio of bot turns that correspond to sub-sequences in the decision tree (with length more than two). A high score indicates that the model is consistent and capable of engaging the user in a conversation according to the tree. A low score indicates that the model diverges from the decision tree without creating the expected series of questions but it also indicates that contrary to a finite state dialog approach where the model is constrained to follow the transitions defined by the finite state automaton, our models can learn new dialog transitions. The observed sequence rates (0.35 and 0.26) suggest both that the models have correctly learned transitions sequences that were defined as natural sounding by the expert and that they can deviate from those, learning new ways to conduct the dialog. We leave a detailed exploration of how these deviations could be used to create alternative dialog paths and thereby enrich the model for further research.

The Acute-Eval results are more nuanced. While the satisfaction (Table 6) and the interest scores (Table 5) are higher for the generative model, the classification model is found more human sounding, more coherent and is preferred for a long conversation. This in lines with previous results (Zhang et al., 2018; Dinan et al., 2018) where retrieval models (approximated here by our hybrid classification/retrieval model) were found to score very well in dialog level evaluations because they return human-written utterances from the training set and thus do not suffer from decoding mistakes present in generative models.

## 7 Conclusion

Using paraphrase identification techniques and a decision tree to model expert knowledge about doctor-patient interactions, we proposed a novel method to create training data for dialog models and we used data created using this method to learn health chatbots that cover the main topics standardly used in the questionnaires of clinical studies. We compared two models, a generative and a hybrid classification/retrieval model and we showed that the expert knowledge captured by the decision tree both helps guide learning and facilitate error analysis.

Results analysis highlights three main directions for future research. First, additional paraphase techniques could be explored to create a more balanced dataset. As shown in Table 3, the quantity of training data available for each domain varies greatly. We are currently exploring whether paraphrase generation (rather than paraphrase extraction) could help address this issue. Second, longer, richer dialogs could be obtained by extending the expert decision tree. Here the American Medical Association Family Medical Guide (Kunz, 1982) may be used to obtain a new dataset with longer and more precise interaction between doctor and patient, giving more advice and information about patients state. Third, even in a clinical study context, human dialogs will often mix open-ended chit-chat with targeted health domain interactions. It would be interesting to extend our approach with strategies that engage the user to talk more about his/her problems e.g., by using ensemble of bots (Papaioannou et al., 2017).

# References

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *ICLR*.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany, August. Association for Computational Linguistics.

Paul Green and Lisa Wei-Haas. 1985. The rapid development of user interfaces: Experience with the wizard of oz method. *Proceedings of the Human Factors Society Annual Meeting*, 29(5):470–474.

Jeffrey R. M. Kunz. 1982. *The American Medical Association Family Medical Guide*. New York: Random House.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Xinzhu Lin, Xiahui He, Qin Chen, Huaixiao Tou, Zhongyu Wei, and Ting Chen. 2019. Enhancing dialogue symptom diagnosis with global attention and symptom graph. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5033–5042, Hong Kong, China, November. Association for Computational Linguistics.

Sourabh Majumdar, Serra Sinem Tekiroglu, and Marco Guerini. 2019. Generating challenge datasets for task-oriented conversational agents through self-play. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 693–702, Varna, Bulgaria, September. INCOMA Ltd.

Ioannis Papaioannou, Amanda Cercas Curry, Jose Part, Igor Shalyminov, Xu Xinnuo, Yanchao Yu, Ondrej Dusek, Verena Rieser, and Oliver Lemon. 2017. Alana: Social dialogue using an ensemble model and a ranker trained on user feedback. In *2017 Alexa Prize Proceedings*.

Alec Radford and Tim Salimans. 2018. Improving Language Understanding by Generative Pre-Training. *OpenAI*, pages 1–12.

Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51, New Orleans - Louisiana, June. Association for Computational Linguistics.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207, Melbourne, Australia, July. Association for Computational Linguistics.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain, April. Association for Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of AAAI*, volume 33.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:2204–2213.