Gathering Information and Engaging the User ComBot : A Task-Based, Serendipitous Dialog Model for Patient-Doctor Interactions

Anna Liednikova, Philippe Jolivet, Alexandre Durand-Salmon, Claire Gardent

Gathering Information and Engaging the User ComBot : A Task-Based, Serendipitous Dialog Model for Patient-Doctor Interactions

Anna Liednikova[&,#], **Philippe Jolivet**[&], **Alexandre Durand-Salmon**[&], **Claire Gardent**[†]
[&] ALIAE
[#] Université de Lorraine
[†] CNRS/LORIA
{philippe.jolivet,alexandre.durand-salmon}@aliae.io
{anna.liednikova,claire.gardent}@loria.fr

Abstract

We focus on dialog models in the context of clinical studies where the goal is to help gather, in addition to the closed set of information collected based on a questionnaire, serendipitous information that is medically relevant. To promote user engagement and address this dual goal (collecting both a predefined set of data points and more informal information about the state of the patients), we introduce an ensemble model made of three bots: a task-based, a follow-up and a social bot. We introduce a generic method for developing follow-up bots. We compare different ensemble configurations and we show that the combination of the three bots (i) provides a better basis for collecting information than just the information seeking bot and (ii) collects information in a more efficient manner that an ensemble model combining the information seeking and the social bot.

1 Introduction

Current work on Human-Machine interaction focuses on three main types of dialogs: task-based, open domain and question answering conversational dialogs. The goal of task-based models is to gather the information needed for a given task e.g., gathering the price, location and type of a restaurant needed to recommend this restaurant. Usually trained on social media data (Roller et al., 2020) (Adiwardana et al.), open domain conversational models aim to mimick open domain conversation between two humans. Finally, question answering conversational models seek to model dialogs where a series of inter-connected questions is asked about a text passage.

In this paper, we consider dialog models in the context of clinical studies i.e., dialog models which are used to collect the information needed by the medical body to assess the impact of the clinical trial on a cohort of patients (e.g., information about their mood, their activity, their sleeping patterns). In the context of these clinical studies, the goal of the dialog model is two-fold. A first goal is to collect a set of pre-defined data points i.e., answers to a set of pre-defined questions specified in a questionnaire. A second goal is to gather relevant serendipitous information i.e., health related information that is not addressed by the questionnaire but that is provided by the user during the interaction and which may be relevant to understand the impact of the therapy investigated by the clinical study. This requires keeping the user engaged and prompting him/her with relevant follow-up questions.

To model these three goals (collecting a predefined set of data points, keeping the user engaged and gathering more informal information about the state of the patient), we introduce an ensemble model which combines three bots: a task-based bot (MEDBOT) whose goal is to collect information about the mood, the daily life, the sleeping pattern, the anxiety level and the leisure activities of the patients; a follow-up bot (FOLLOWUPBOT) designed to extend the task-based exchanges with health-related, follow-up questions based on the user input; and an empathy bot (EMPATHYBOT) whose task is to reinforce the patient engagement by providing empathetic and socially driven feedback.

Our work makes the following contributions.

- We introduce a model where interactions are driven by three main goals: maintaining user engagement, gathering a predefined set of information units and encouraging domain related user input.
- We provide a generic method to create training data for a bot that can follow-up on the user response while remaining in a given domain (in this case the health domain).
- We show that such a follow-up bot is crucial to support both information gathering and user

Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations, pages 21–29

engagement and we provide a detailed analysis of how the three bots interact.

2 Related Work

Several approaches have explored the use of ensemble models for dialog. While Song et al. (2016) proposed an ensemble model for human-machine dialog which combines a generative and a retrieval model, further ensemble models for dialog have focused on combining agents/bots designed to model different conversation strategies. Yu et al. (2016) focus on open domain conversation and combines three agents, two to improve dialog coherence (ensuring that pronouns can be resolved and maximising semantic similarity with the current context) and one to handle topic switch (moving to a new topic when the retrieval confidence score is low). The ALANA ensemble model (Papaioannou et al., 2017b,a), developed for the Amazon Alexa Challenge i.e., for open domain chitchat, combines domain specific bots used to provide information from different sources with social bots to smooth the interactions (by asking for clarification, expressing personal views or handling profanities). Similarly, Yu et al. (2017) introduces a dialog model which interleaves a social and a task-based bot. Conversely, Gunson et al. (2020) showed that success of interleaving depends on the context and that in a public setting, users either prefer purely task-based systems or fail to see a difference between task-based and a richer ensemble model combining task-based and social bots.

Our work differs from these previous approaches in that we combine a standard, task-based model with both a social bot and a domain specific, follow-up bot. This allows both for more natural dialogs (by following up on the user input rather than systematically asking about an item in the predefined set of topics) and for additional relevant, health related information to be gathered.

3 ComBot, an ensemble Model for Repeated Task-Based Interactions

We introduce the three bots making up our ensemble model and the ensemble model combining them.

3.1 Medical Bot

MEDBOT is a retrieval model which uses the pretrained ConveRT dialog response selection model (Henderson et al., 2019) to retrieve a query from the MedTree Corpus (Liednikova et al., 2020). It is designed to collect information from the user based on a predefined set of questions contained in a questionaire.

The MedTree Dataset. The MedTree corpus (Liednikova et al., 2020) was developed to train a task-based, information seeking, health bot on five domains: sleep, mood, anxiety, daily tasks and leisure activities. It was derived from a dialog tree provided by a domain expert (i.e., a physician) and designed to formalise typical patient-doctor interactions occurring in the context of a clinical study. In that tree, each branch captures a sequence of (Doctor Question, Patient Answer) pairs and each domain is modeled by a separate tree with the root introducing the conversation (initial question) and the leaves providing a closing statement. The MedTree corpus is then derived from this tree by extracting from each branch of the tree, all context-question pairs, where the context consists of a sequence of patient-doctor-patient turns present on that branch and the question is the following doctor question. A fragment of the decision tree created for the sleep domain and an example dialog are shown in Figure 1.

There are two versions of the MedTree corpus: one consisting of only the context/question pairs derived from the dialog tree (INIT) and the other including variants of these pairs based on paraphrases extracted from forum data (ALL). In (Liednikova et al., 2020), the ALL corpus is used to train a generative and a classification model. In our work, we use (a slightly modified version[1] of) the INIT corpus instead, as its small size facilitates retrieval (the number of candidates is small) and preliminary experimentations showed better results when using the INIT corpus.

Model. ConverT is a Transformer-based Encoder-Decoder which is trained on Reddit (727M input-response pairs) to identify the dialog context most similar to the current context and to retrieve the dialog turn following this context. In order to retrieve from the MedTree corpus, the question that best fits the current dialog context, the MEDBOT model compares the last three turns of the current dialog with contexts from the MedTree Corpus. The model identifies the

[1]The modifications consists in shortening the questions, changing all leaves to statements and adding meta-statements about the dialog to account for cases where the user indicates misunderstanding or agreement

MedTree corpus context with the highest similarity score[2] and outputs the question following that context. If the selected question has already been asked in the dialog generated so far and provided it is not a question such as "What other things would you like to share with me ?", we retrieve the next best question that is not a repetition. No fine-tuning is done due to the small amount of data.

3.2 Follow-Up Bot

One main motivation behind the use of a health-bot in clinical studies is to complement the information traditionally gathered through a fixed questionnaire filled in each week by the patients with serendipitous information i.e., information that is not actively queried by the questionnaire but that is useful to analyse the cohort results.

The MEDBOT model introduced in the previous section is constrained to address only those topics which are present in the dialog tree, in effect, modeling a closed questionnaire. To allow for the collection of serendipitous health information, we develop the FOLLOWUPBOT whose function is to generate health-related questions which are not predicted by the dialog tree but which naturally follow from the user input. The main difference of FOLLOWUPBOT from MEDBOT is the way it retrieves questions that are not in the sequence, but the ones that occurs in the same context even if the question itself doesn't share the lexions with the previous turns. Rather than artificially restricting the dialog to the limited set of topics pre-defined by the dialog tree, the combined model (MEDBOT + FOLLOWUPBOT) allows for transitions based either on the dialog tree or on health-related, follow-up questions. In that sense, FOLLOWUPBOT allows not only for the collection of health-related serendipitous information but also for smoother dialog transitions.

Like MEDBOT, FOLLOWUPBOT used the pre-trained ConveRT model to retrieve context appropriate queries from a dialog dataset. In this case however, the queries are retrieved from the HealthBoard dataset, a new dataset we created to support follow-up questions in the health domain.

The Healthboard Dataset. This dataset consists of (s, q) pairs where s is a (health related) statement and q is a follow-up question for that statement. We extract this dataset from the Healthboard forum [3] as follows. We first select 16 forum categories (listed in Table 1) that are relevant to our five domains. In the forum, each category includes multiple conversational threads, each thread consists of multiple posts and each post is a text of several paragraphs that can be split into sentences. In total, we collect 175,789 posts from 31,042 threads with 5.68 posts in average per thread. We then segment each post into sentences using the default NLTK sentence segmenter. We label each sentence with a dialogue act classifier in order to distinguish statements ("sd" label) from questions ("qo" label). For this labelling, we fine-tune the Distilbert Transformer-based classification model [4] on the Switchboard Corpus Stolcke et al. (2000) using 6 classes "qo" (Open-Question), "sd" (Statement-non-opinion), "ft" (Thanking), "aa" (Agree/Accept), "%" (Uninterpretable) and "ba" (Appreciation). For each question q (i.e., sentence labelled "qo") in each thread T, we gather all statements (i.e., all sentences labeled as "sd") which precede q in T into a pool of candidate statements[5]. As dialogue turns in bots should remain short, we filter sentences that have more than 100 tokens. For each candidate statement, we calculate its similarity with the question using the dot product on their ConveRT embeddings. We filter out all candidate statements whose score with the question is less than 0.6. If after filtering the resulting pool contains at least one candidate, we select the top-ranked statement and add the statement-question pair pair to the dataset. The resulting dataset contains 3,181 (statement, question) pairs.

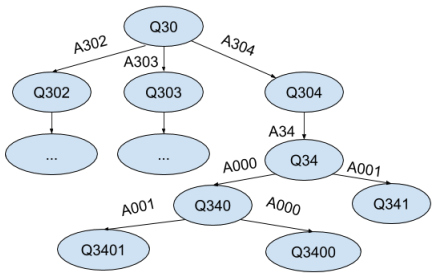
Model. Similar to the MEDBOT model, the FOLLOWUPBOT model used the pre-trained ConveRT model to compare the current dialog context (the preceding three turns) with the statements contained in the HealthBoard dataset using the inner product. The top-20 candidates are then retrieved and filtered using Maximal Marginal Rel-

[2]Both contexts are encoded using ConveRT as average of embeddings of the last turn and concatenation of preceding ones. The inner product is used to compute similarity.

[3]https://www.healthboards.com/

[4]https://huggingface.co/distilbert-base-uncased

[5]We do not restrict the set of candidates at that stage i.e., we consider all posts that precede the question within the question thread and all statements in these posts no matter how far away the statement is from the question. In practice, the set of such statements has limited size and distance does not seem to matter too much although an investigation of that factor would be interesting. We leave this question open for further research as it is not central to our paper.



<Q30> <tree_pos> What is the most difficult for you about your sleep ?
<A304> <tree_pos> I wake up early in the morning
<Q304> <tree_pos> What time did you wake up today ?
<A34> <tree_pos> I woke up at 5 am
<Q34> <tree_pos> Do you wake up spontaneously but without enough sleep to feel rested all day long ?
<A000> <tree_pos> no
<Q340> <tree_pos> Could you describe how do you wake up ?
<A001> <tree_pos> yes, the alarm clock wakes me up
<Q3401> <tree_pos> Thanks to these informations, your physician will have more information at his disposal for your next visit

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

Category	Threads	Posts	Avg
anxiety	6852	38523	5.63
anxiety tips	42	71	1.69
chronic fatigue	670	3856	5.77
chronic pain	646	4893	7.59
depression	5327	32998	6.21
depression tips	27	51	1.89
exercise fitness	1583	8142	5.16
general health	7279	29858	4.11
healthy lifestyle	104	621	5.97
pain management	4985	38738	7.79
panic disorders	1314	8376	6.39
share your anxiety story	42	42	1
share your depression story	55	71	1.29
share your pain story	28	42	1.50
sleep disorders	1671	7656	4.59
stress	415	1973	4.76

Table 1: Forum Categories used for the Creation of the HealthBoard Dataset

evance (MMR) (Carbonell and Goldstein, 1998) with $\lambda = 0.5$ to control for repetitions[6]. Next, we compute the similarity between the remaining selected questions and the questions included in the current dialog context (all preceding dialog turns) and we exclude candidates with similarity score 0.8 or higher. After filtering, the top ranking candidate is selected and the associated follow-up question is output.

3.3 Empathy Bot

As the name suggests, the role of the EMPATHYBOT is to engage the user by showing empathy. For this bot, we use Roller et al. (2020) generative model which was pre-trained on a variant of Reddit discussion (Baumgartner et al., 2020) and fine-tuned on the ConvAI2 (Zhang et al., 2018), Wizard of Wikipedia (Dinan et al., 2019), Empathetic Dialogues (Rashkin et al., 2019), and Blended Skill Talk datasets (BST) (Smith et al., 2020) to optimize engagigness and humanness in open-domain conversation.

3.4 Ensemble Model (ComBot)

Each bot provides a single candidate. To rank them, we encode the whole current dialog context and each candidate response using the ConveRT encoder, we calculate similarity (dot product) for each candidate/context pair and we select the candidate with highest similarity score. In case all candidates scores are less than 0.1, we consider that there is no good response and we end the conversation.

4 Experiments

4.1 Data

Table 2 shows some statistics for the corpora used for pretraining (ConveRT, Blender) and for retrieval (INIT, HealthBoard). For MEDBOT and FOLLOWUPBOT, we use the ConveRT model from PolyAI [7]. For EMPATHYBOT, we use the Blender model with 90M parameters from the ParlAI library [8].

One benefit of the ensemble approach is that several models can be combined, each modelling different types of dialog requirements. We compare different configurations of our three bots: COMBOT (which combines the three bots), MEDBOT (using only the task-based bot), MED+EMPATHYBOT an ensemble model which combines the task-based (MEDBOT) and the social bot (EMPATHYBOT) and MEDBOT+ FOLLOWUPBOT, a bot combining the task-based and the follow-up question bot.

We first use automatic metrics and global satisfaction scores to compare the four models. We restrict the Acute-Eval, human-based model comparison to the two best performing systems namely,

[6]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 a selected question). A λ value of 0.5 favors similarity and diversity equally, both matter equally.

[7]https://github.com/connorbrinton/polyai-models/releases/tag/v1.0

[8]https://parl.ai/projects/recipes/

COMBOT and MEDBOT.

4.2 Evaluation

As there does not exist a dataset of well-formed health-related dialogs whose aim is both to answer a clinical study questionaire and to allow for serendipitous interactions, we have no test set on which to compare the output of our dialog models. Moreover, as has been repeatedly argued, reference-based, automatic metrics such as BLEU or METEOR, fail to do justice to the fact that a dialog context usually has many possible continuations. We therefore use reference-free automatic metrics and human assessment for evaluation.

Human evaluation. We use the MTurk platform to collect human-bot dialogs for our four models (COMBOT, MEDBOT and MED+EMPATHYBOT) and ask the crowdworkers to provide a satisfaction rate at the end of their interaction with the bot. We then run a second MTurk crowdsourcing task to grade and compare dialogs pairs produced by different models.

To collect dialogs, we ask participants to interact with the bot for as long as they want. The conversation starts randomly with one of the initial questions of MEDBOT. The interaction stops either when all candidates scores are less than 0.1 (cf. Section 3.4) or when the user ends the conversation. For each model, we collect 50 dialogs. Each annotator interacts at most once with a bot.

At the end of each human-bot conversation, the annotator is asked to rate satisfaction on a 1-5 Likert scale (a higher score indicates more satisfaction).

Assigning a satisfaction score to a single dialog is a highly subjective task however with scores suffering from different bias and variance per annotators (Kulikov et al., 2019). As argued by Li et al. (2019), comparing two dialogs, each produced by different models, and deciding on which dialog is best with respect to a predefined set of questions, helps support a more objective evaluation. We therefore use the Acute-Eval human evaluation framework to compare the dialogs collected using different bots. Since the automatic evaluation (cf. Section 5.1) shows that COMBOT and MEDBOT are the best systems, we compare only these two systems asking annotators to read pairs of dialogs created by these two bots and to then answer the pre-defined set of questions recommended by Li et al. (2019)'s evaluation protocol namely:

- 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?

We report the percentage of time one model was chosen over the other.

For this comparison, we consider 50 dialog pairs (one dialog produced by COMBOT, the other by MEDBOT) and for each Acute-Eval question, collected 50 judgments, one per dialog pair. We had ten annotators, each annotating at most 5 dialog pairs. To maximise similarity between the dialogs being compared, we create the dialog pairs by computing euclidean distance between context embeddings of MEDBOT and COMBOT dialogue sets. Then we composed a pair of two closest items and excluded them from the choice in the next iteration.

Automatic Metrics. After collecting dialogues we perform their automatic evaluation. All scores are computed on the 50 bot-human dialogs collected for a given model. Table 3 shows the result scores averaged over 50 dialogs.

To measure *coherence*, we exploit the unsupervised model CoSim introduced by Mesgar et al. (2019); Xu et al. (2018); Zhang et al. (2017). This model measures the coherence of a dialog as the average of the cosine similarities between ConveRT embedding vectors of its adjacent turns.

To assess *task success*, we count the number of unique medical entities (Slots) mentioned. We do this using the clinical NER-model from gthe Stanza library (Zhang et al., 2020) [9], a model trained on the 2010 i2b2/VA dataset (Uzuner et al., 2011) to extract named entities denoting a medical problem, test or treatment. We report the average number of medical entities both per dialog and in the user turns (to assess how much medical information comes from the user).

Following Yu et al. (2017), we also calculate *Information gain (InfoGain)*, the average number of unique tokens per dialog and *Conversation Length (ConvLen)*, the average number of turns in the overall dialog.

[9] http://stanza.run/bio

	Reddit	ConvAI2	WoW	EmpaDial	BSD	INIT	HealthBoard
Nb of context-question pairs		211803	83011	76673	27018	168	3181
Nb of distinct turns	1.50B	267945	165213	88757	53335	154	73140
Nb of tokens	568B	3791971	2720426	2625338	912857	3688	202389
Nb of tokens per turn (Avg, Max, Min)		8.95	16.39	17.12	16.89	6.92	11.5
Vocabulary size		20707	95590	59438	52561	306	7321

Table 2: Corpus statistics (Reddit: pre-training corpus for ConveRT and the Empathy bot. ConvAI2, WoW, EmpaDial and BSD: Datasets used to fine-tune the Empathy Bot. INIT: used for the MedBot retrieval step. HealthBoard: for FollowUp Bot Fine-Tuning and Retrieval .)

Model	Satisf.	CoSim	Slots	ConvLen	InfoGain	UserQ
MEDBOT	3.94	0.26	6.24 (1.68)	28.46	108.82 (3.82)	0.08 (4)
MEDBOT+ FOLLOWUPBOT	3.18	0.34	11.65 (3.22)	36.06	153.23 (4.25)	0.47 (23)
MEDBOT+ EMPATHYBOT	3.77	0.34	3.87 (1.46)	30.29	140.19 (4.63)	0.68 (33)
COMBOT	3.72	0.36	7.12 (2.82)	21.96	124.82 (5.68)	0.48 (24)

Table 3: Satisfaction Scores (Satisf.) and Results of the Automatic Evaluation. CoSim: Average Cosine Similarity between adjacent turns. Slots: Average Number of Medical Entities per dialogue (in brackets: average number in the user turns). ConvLen: Average Number of turns per dialog. InfoGain: Average number of unique tokens per dialog (in brackets: normalised by dialog length). UserQ: number of questions asked by Human (in bracket: total number for 50 dialogs). All metrics are averaged over the 50 Human-Bot dialogs collected for each model.

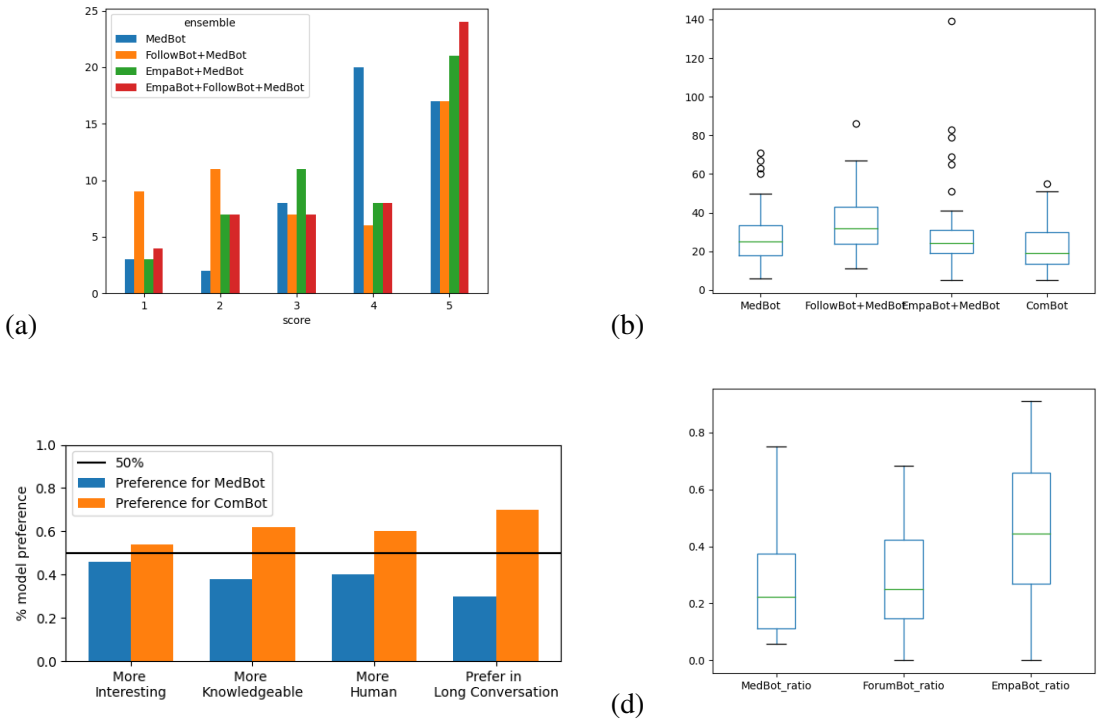


Figure 2: (a) Distribution of the Satisfaction Scores for each configuration, (b) Conversation length distribution for MedBot and ComBot, (c) Acute-Eval results for both systems, (d) Majority bot ratio in COMBOT

Finally, we compute the number of questions asked by the user (*UserQ)* as an indication of the user trust and engagement. We compute both the total number of questions present in the 50 dialog collected for a given model and the average number of question per dialog.

5 Results and Discussion

We compare four models using automatic metric and absolute satisfaction scores. Based on this first evaluation, we compare two of these models using the Acute-Eval human evaluation framework. We display an example dialog and discuss the respec-

tive use of each bot in the COMBOT model.

5.1 Automatic Evaluation and Absolute Satisfaction Scores

Table 3 shows the absolute satisfaction scores (i.e., scores provided on the basis of a single dialog rather than by comparing dialogs produced by different models) and the results of the automatic evaluation for the four models mentioned above.

ComBot provides a better basis for collecting information than MedBot. The automatic scores show that COMBOT consistently outperforms MEDBOT on informativity (Slots, InfoGrain) while allowing for shorter dialogs (ConvLen). In other words, COMBOT allows for a larger range of informational units (words and medical named entities) to be discussed in fewer turns.

ComBot collects information in a more user-friendly, more efficient manner than Med+EmpathyBot. While the InfoGain scores are higher for MED+EMPATHYBOT and MEDBOT+FOLLOWUPBOT than for COMBOT (InfoGain: 140.19 and 153.23 vs. 124.82), this is achieved at the cost of much longer dialogs (ConvLen: 30.29 and 36.06 vs. 21.96; cf. also Figure 2b) In fact, when normalising InfoGain by the number of dialog turns (ConvLen), we see that in average, a turn in COMBOT dialogs contains a much higher number of unique tokens (i.e., is more informative) than for MEDBOT (3.82), MEDBOT+EMPATHYBOT (4.63) or MEDBOT+FOLLOWUPBOT (4.25).

ComBot allows for more coherent dialogs. In terms of quality, the differences in satisfaction scores between the three models is not statistically significant ($p < 0.05$, T-test). For dialog coherence (Measured by CoSim) however, COMBOT scores highest (0.36) and the difference with MEDBOT is statistically significant ($p < 0.05$, T-test). This suggests that follow up questions help support smoother transitions between dialog turns.

5.2 Comparative Human Evaluation

The results of the comparative human evaluation are presented in Figure 2.

ComBot is judged more knowledgeable, more interesting, more human and better for long conversations. COMBOT outperforms MEDBOT on all Acute-Eval questions (Figure 2c).

In particular, users find COMBOT more knowledgeable by a large margin. This is in line with the automatic metrics results (higher COMBOT values for Slots and InfoGain) and is likely due to the fact that the COMBOT model supports the use of health-related, follow-up questions which in turn allows for a wider range of medical issues to be discussed than just those present in the MEDBOT corpus.

Users also show a clear preference for COMBOT in long conversations (Figure 2a). While this seems to contradict the fact that both models have similar satisfaction score, we conjecture that the high MEDBOT satisfaction score is an artefact of the MEDBOT model. Since the MEDBOT coverage is restricted, the users have low expectations and correspondingly give high satisfaction scores (they are easily satisfied because their expectations are low). An indication of these low user expectations is given by the number of questions asked : when users feel that the system they interact with is unrestricted, they will feel comfortable asking questions and will start to do so. Conversely, if they feel the model is restricted, they will refrain from asking questions. The results show a much higher number of questions for users interacting with COMBOT (Table 3)

5.3 Component analysis

Figure 3 displays an example Human-Bot dialog using the COMBOT model which illustrates the interactions between the three composing bots: the EMPATHYBOT closes the conversation with social chit-chat, the FOLLOWUPBOT responds to the user turn and MEDBOT asks questions from the dialog tree whenever suitable.

The proportion of turns generated by each bot (cf. Figure 2d) varies from one dialog to another, illustrating the capacity of the ensemble model to adapt to various dialog users and contexts. We find that in 55% of the collected dialogs, a majority of turns (i.e., more than 33% of the turns) is generated by the EMPATHYBOT model; in 29% of the cases by the FOLLOWUPBOT and in 16% of the cases by the MEDBOT[10]

We also observe interesting dependencies and correlations. MEDBOT is triggered twice more of-

[10]Since a COMBOT dialog has an average of 21 turns and only half of those are generated by the bot, this means that for 55% of the collected dialogs, the dialog contains more than 3 "social" dialog turns (turns generated by EMPATHYBOT). Similarly, 29% of the collected dialogs contain more than 3 follow-up turns (FOLLOWUPBOT) and 16% more than 3 task-based turns (MEDBOT).

BOT: : *[opening question]* What would you like to share today?
USER: : i feel soo tired
BOT: : /FOLLOWUPBOT/ Is something making you hyper at night?
USER: : maybe i am so stressed that i cannot sleep
BOT: : /MEDBOT/ What wakes you up at night ?
USER: : nothing, i just cannot fall asleep
BOT: : /FOLLOWUPBOT/ Are your sleep issues anxiety related?
USER: : i think so
BOT: : /FOLLOWUPBOT/ Also do you snore when you do sleep possibly sleep apnea?
USER: : i do not know
BOT: : /MEDBOT/ Do you feel anxiety or stress during nights awakenings ?
USER: : yes, a bit
BOT: : /MEDBOT/ Anxiety and sleep may be connected, it is time to speak with your physician
USER: : okay
BOT: : /FOLLOWUPBOT/ What about getting a second opinion from another doctor about your situation?.
USER: : why not
BOT: : /FOLLOWUPBOT/ What does your doctor say about coming off of them and then trying to go back on.
USER: : i do not know
BOT: : /MEDBOT/ Are you not able to fall asleep regularly or punctually ?
USER: : regularly
BOT: : /EMPATHYBOT/ i hope things get better for you , and that you are able to enjoy your time with your family .
USER: : thank you
BOT: : /EMPATHYBOT/ you are welcome , and i wish you the best of luck with the rest of your life . :)

Figure 3: Example Human-ComBot dialog

ten after FOLLOWUPBOT (30 cases) than after EMPATHYBOT (12 cases) – this indicates that follow-up questions help bringing the user back to the questions contained in the dialog tree.

6 Conclusion

A qualitative analysis of the collected dialogs indicates several directions for further research.

Negation is often not recognised leading to interactions in which the model continues discussing a topic which was declared as irrelevant by the user. Another difficulty is knowing when to end the conversation. Long ones are good to complete the task, but bad for people who are ready to finish conversation but feel forced to continue. To improve user engagement, a possibility would be to explore whether the information provided by sentiment analysers could be exploited to help maintain a positive interaction. By detecting polarity, it could also help improve negation handling.

Another key issue concerns the emotional impact of the dialog on the user. An interaction with the bot might highlight a health issue the user was not aware of resulting in increased user stress. In such a situation, a good policy would be to provide the user with some notion of solution, some piece of information or advice which can help her face the situation and if possible, incite her to act to improve her health. Indeed some of the dialogs collected with COMBOT show that users sometimes ask for help. Here a knowledge-based agent could be useful either to provide facts that are related to the topic at hand or to highlight the connections between facts that have been mentioned in the dialog.

Acknowledgements

We thank the anonymous reviewers for their feedback. We would like to acknowledge Farnaz Ghassemi for her help in developing the FOLLOWUPBOT. We gratefully acknowledge the support of the ALIAE company, the French National Center for Scientific Research, and the ANALGESIA Institute Foundation.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu Quoc, and V Le. Towards a Human-like Open-Domain Chatbot. Technical report.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset.

Jaime Carbonell and Jade Goldstein. 1998. Use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, pages 335–336, New York, New York, USA. ACM Press.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard

of wikipedia: Knowledge-powered conversational agents.

Nancie Gunson, Weronika Sieińska, Christopher Walsh, Christian Dondrup, and Oliver Lemon. 2020. It's good to chat? evaluation and design guidelines for combining open-domain social conversation with task-based dialogue in intelligent buildings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20, New York, NY, USA. Association for Computing Machinery.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2019. Convert: Efficient and accurate conversational representations from transformers.

Ilia Kulikov, Alexander H. Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling.

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*.

Anna Liednikova, Philippe Jolivet, Alexandre Durand-Salmon, and Claire Gardent. 2020. Learning healthbots from training data that was automatically created using paraphrase detection and expert knowledge. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Mohsen Mesgar, Sebastian B ¨ Ucker, and Iryna Gurevych. 2019. A Neural Model for Dialogue Coherence Assessment. Technical report.

Ioannis Papaioannou, Amanda Cercas Curry, Jose Part, Igor Shalyminov, Xu Xinnuo, Yanchao Yu, Ondrej Dusek, Verena Rieser, and Oliver Lemon. 2017a. Alana: Social Dialogue using an Ensemble Model and a Ranker trained on User Feedback. In *2017 Alexa Prize Proceedings*.

Ioannis Papaioannou, Amanda Cercas Curry, Jose L Part, Igor Shalyminov, Xinnuo Xu, Yanchao Yu, Ondřej Dušek, Verena Rieser, and Oliver Lemon. 2017b. An ensemble model with ranking for social dialogue. *arXiv preprint arXiv:1712.07558*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills.

Yiping Song, Rui Yan, Xiang Li, Dongyan Zhao, and Ming Zhang. 2016. Two are Better than One: An Ensemble of Retrieval- and Generation-Based Dialog Systems.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Ö. Uzuner, B.R. South, S. Shen, and S.L. DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Zhou Yu, Alan W. Black, and Alexander I. Rudnicky. 2017. Learning conversational systems that interleave task and non-task content. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4214–4220. AAAI Press.

Zhou Yu, Ziyu Xu, Alan W Black, and Alexander Rudnicky. 2016. Strategy and policy learning for non-task-oriented conversational systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 404–412, Los Angeles. Association for Computational Linguistics.

Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2017. Reinforcing Coherence for Sequence to Sequence Model in Dialogue Generation. Technical report.

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? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D. Manning, and Curtis P. Langlotz. 2020. Biomedical and Clinical English Model Packages in the Stanza Python NLP Library.