Would you like to tell me more? Generating a corpus of psychotherapy dialogues

Seyed Mahed Mousavi[1], Alessandra Cervone[2]*, Morena Danieli[1], Giuseppe Riccardi[1]
[1]Signals and Interactive Systems Lab, University of Trento, Italy
[2]Amazon Alexa AI
{mahed.mousavi, giuseppe.riccardi}@unitn.it

Abstract

The acquisition of a dialogue corpus is a key step in the process of training a dialogue model. In this context, corpora acquisitions have been designed either for open-domain information retrieval or slot-filling (e.g. restaurant booking) tasks. However, there has been scarce research in the problem of collecting personal conversations with users over a long period of time. In this paper we focus on the types of dialogues that are required for mental health applications. One of these types is the follow-up dialogue that a psychotherapist would initiate in reviewing the progress of a Cognitive Behavioral Therapy (CBT) intervention. The elicitation of the dialogues is achieved through textual stimuli presented to dialogue writers. We propose an automatic algorithm that generates textual stimuli from personal narratives collected during psychotherapy interventions. The automatically generated stimuli are presented as a seed to dialogue writers following principled guidelines. We analyze the linguistic quality of the collected corpus and compare the performances of psychotherapists and non-expert dialogue writers. Moreover, we report the human evaluation of a corpus-based response-selection model.

1 Introduction

The idea of developing conversational agents as Personal Healthcare Agents (PHA) (Riccardi, 2014) has gained growing attention in recent years for various domains including mental health (Fitzpatrick et al., 2017; Abd-alrazaq et al., 2019; Ali et al., 2020). Most of the conversational agents in the mental health domain are created using rule-based and simple predefined tree-based dialogue flows, resulting in limited understanding of the user input and repetitive responses by the agent. These limitations lead to shallow conversations and weak user engagement (Abd-Alrazaq et al., 2021).

The major reasons for such limitations are the complexity of conversations, the lack of dialogue data and domain knowledge. The conversations about mental state issues are very complex because they usually encompass personal feelings, user-specific situations, different spaces of entities, and emotions. In this domain, the state-of-the-art data-driven frameworks are not applicable and domain knowledge is very scarce. The two main approaches to collect dialogue data for the purpose of developing data-driven dialogue agents are either acquiring user interaction data via user simulators and hand-designed policies (Li et al., 2016), or to collect large sets of human-human conversations in different user-agnostic settings (Budzianowski et al., 2018; Gopalakrishnan et al., 2019; Zhang et al., 2018). These approaches have been used for goal-oriented agents (e.g. reservations of restaurants) or open-domain agents answering questions about a finite set of topics (e.g. news, music, weather, games etc.). However, neither of the above approaches can address the need for personal conversations which include user-specific recollections of events, objects, entities and their relations. Last but not least, state-of-the-art conversational agents cannot carry out engaging and appropriate single-user multi-session conversations. However, personal conversations' requirements include the ability of carrying out multi-session conversations over several weeks or months.

In this paper, we propose a novel methodology to collect corpora of follow-up dialogues for the mental health domain (or domains with the similar characteristics). Psychotherapists deliver interventions over a long period of time and need to monitor or react to patients' input. In this domain, dialogue follow-ups are a critical resource for psychotherapists to learn about the life events of the narrator as well as his/her corresponding thoughts and emotions in a timely manner. In Figure 1 we describe the proposed workflow for the

*The work was done while at the University of Trento, prior to joining Amazon Alexa AI.

Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations, pages 1–9

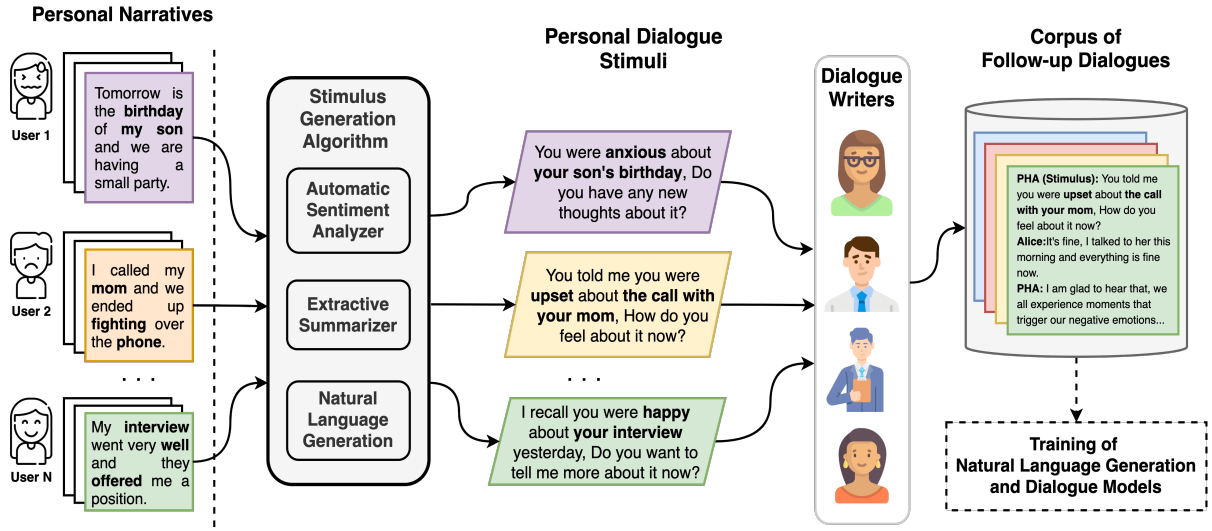

Figure 1: The workflow for the elicitation of follow-up dialogues starting from the personal narratives collected during psychotherapy (left-hand side) interventions. The stimulus generation algorithm creates a textual stimulus from personal narratives as a seed to dialogue writers. Dialogue writers use the textual stimulus and principled guidelines to generate the follow-up dialogues (right-hand side). The dialogue follow-ups may be used to train dialogue models, response-selection models and natural language generators.

acquisition of personal dialogue data aimed at training dialogue models. We first collect a dataset of personal narratives written by the users who are receiving Cognitive Behavioral Therapy (CBT) to handle their personal distress more effectively[1]. In the next step, the narratives are used to generate stimuli for the follow-up conversations with an automatic algorithm. The first part of the stimulus, the common-ground statement, contains the summary of the narrative the user has previously left and the associated emotions and the second part is a follow-up question aimed at reviewing the users life events. In the last step, the stimuli are presented to writers and they are asked to generate a conversation based on the provided stimulus by impersonating themselves as both sides of the conversation, an approach introduced firstly by Krause et al. (2017), where in our setting the sides are the PHA and the patient.

The main contributions of this paper can be summarized as follows:

- We present a methodology for data collection and elicitation of follow-up dialogues in the mental health domain.
- We present an algorithm for automatically generating conversation stimuli for follow-up dialogues in the mental health domain from a sequence of personal narratives and recollections, with a similar structure that psychotherapists use when reviewing the progress with the patient.
- We evaluate the collected dialogue corpus in terms of the quality of the obtained data, as well as the impact of domain expertise on writing the follow-up dialogues.
- We investigate the suitability of the collected corpus for developing conversational agents in the mental health domain by automatic and human evaluation of a baseline response-selection model.

2 Literature Review

Knowledge grounded dialogue corpora Previously published research have addressed the problem of collecting dialogue data starting from world knowledge facts or predefined persona descriptions. In this regard, Zhang et al. (2018) collected a dataset of conversations conditioned on synthetic persona descriptions for each side of the dialogue using Amazon Mechanical Turk (AMT) workers. Gopalakrishnan et al. (2019) collected a dataset of dialogues grounded in world knowledge by pairing AMT workers to have a conversation based on selected reading sets from Wikipedia and The Washington Post over various topics. Furthermore, Rashkin et al. (2019) have crowdsourced a dataset

[1]This data collection has been approved by the Ethical Committee of the University of Trento.

of conversations with implied user feelings in the context, using AMT workers where a worker writes a personal situation associated to an emotion and in the next step is paired with another worker to have a conversation about the mentioned situation. While useful for chitchat and open-domain conversations, unfortunately these resources are not a good fit to address the needs of the mental health support domain.

Mental health support dialogue corpora The research in this domain is very recent and resources are scarce. "Counseling and Psychotherapy Transcripts" published by Alexander Street Press[2] is a dataset of 4000 therapy session transcriptions on various topics, used as a resource for therapists-in-training. Pérez-Rosas et al. (2016) collected a dataset of 277 Motivational Interviewing (MI) session videos and obtained the transcriptions for each session either directly from the data source, or by recruiting AMT workers. Guntakandla and Nielsen (2018) conducted a data collection process of therapeutic dialogues in Wizard of Oz manner where the therapists were impersonating a Personal Healthcare Agent. The authors recorded 324 sessions of therapeutic dialogues which were then manually transcribed. Furthermore, in the physical health coaching domain, Gupta et al. (2020) collected a dataset of conversations where the expert impersonates a PHA that engages the users into a healthier life style. For this purpose, a certified health coach interacted with 28 patients using a messaging application.

3 Dialogue Follow-Up Generation Methodology

The type of dialogues that we aim at obtaining is different from what has been reported in the literature. While previous works attempted to collect in-the-field therapeutic interactions and convert them into dialogue datasets, we present an elicitation methodology to generate a dataset of follow-up dialogues in the mental health domain, grounded in the personal narratives and with the same conversational structure that the psychotherapists use in order to review the events with the patients in a timely manner.

3.1 Collection of Personal Narratives

A group of 20 Italian native speakers who were receiving Cognitive Behavioral Therapy (CBT) were asked to write notes about the daily events that activated their emotional state. CBT is a psychotherapy technique based on the intuition that it is not the events that directly generate certain emotions but how these events are cognitively processed and evaluated and how irrational or dysfunctional beliefs influence this process (Oltean et al., 2017). A technique commonly used in CBT treatment is the ABC (**A**ntecedent, **B**elief, **C**onsequences). In this technique, the psychotherapist tends to identify the event that has caused the patient a certain emotion by a set of questions to define **A**) what, when and where the event happened, **B**) the patient's thoughts and beliefs about the event and **C**) the emotion the patient has experienced regarding the event. Once dysfunctional thoughts are identified, the patient is guided on how to change them or find more rational and/or functional thoughts (Sarracino et al., 2017).

We recruited 20 users who would meet with their human psychotherapists one session a week and asked them to write notes about the day-life events that caused them an emotional arousal between one

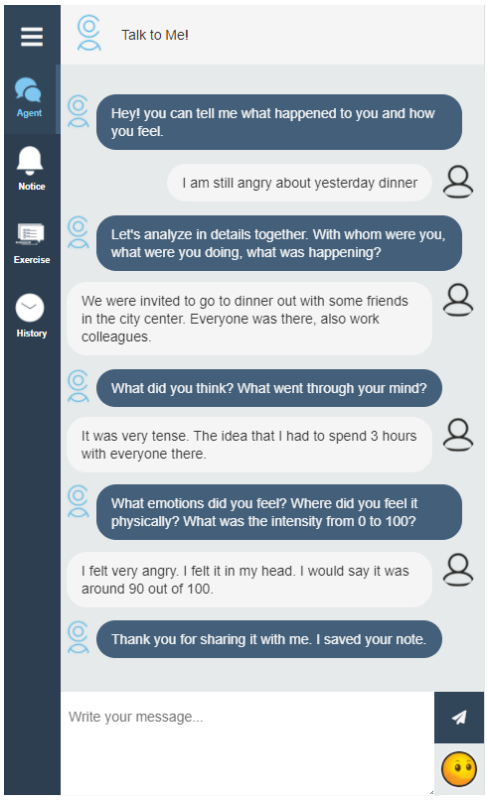

Figure 2: The user interface of the mobile application designed for collecting personal narratives (English translations). The patients were asked to describe events, persons, situations that explained their emotional arousal while answering the ABC questions designed by psychotherapists.

[2] https://alexanderstreet.com/

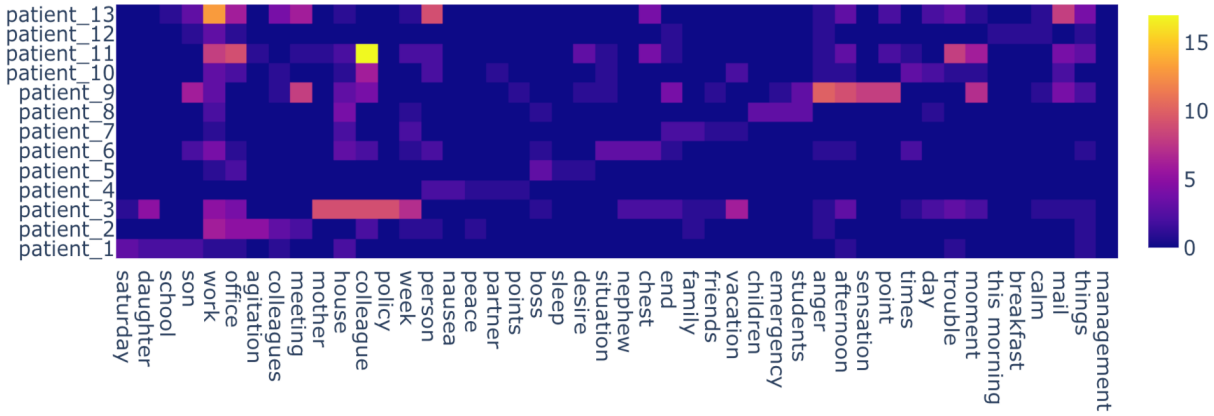

Figure 3: The heat-map of frequent nouns used by the patients in collected personal narratives (English translations). The x-axis represents the nouns extracted from the 5-most frequent list used by each user while the y-axis and z-axis represent the users and the noun frequency, respectively.

session and the following one. For this purpose, a mobile application was designed that the users could interact with for a period of three months, to answer the questions designed by the psychotherapists for the ABC technique, and assign an emotion to the note if possible. The emotions could be selected from a predefined set, equal for all users, including the six basic emotions used in psychological experiments (Happiness, Anger, Sadness, Fear, Disgust and Surprise) (Ekman, 1992), and two other complex emotional states (Embarrassment and Shame) that were considered relevant for this setting. Figure 2 shows the user interface of the application designed for this purpose.

By the end of this step, 224 ABC notes were obtained from 20 users of which 92 notes (written by 13 different subjects) are complete, i.e. the users has answered all the questions completely, and are selected for the generation of the stimuli. Considering the fact that each note, that is the answers to the ABC questions, is about a unique real-life event, we concatenate the answers in each note under the psychotherapists' supervision to convert the notes into personal narratives of one piece. Out of the 92 complete narratives, 18 narratives are assigned an emotion by the user, and 74 notes are not labeled by any emotions. A lexicon-based sentiment analyzer developed by The OpeNER project[3] is used to detect the polarity of the 74 narratives without any expressed emotions, which labeled 61 narratives as either negative or positive and 13 of them as neutral.

Lexical analysis on the selected narratives demonstrates that the language and vocabulary used in the narratives are user-specific. Figure 3 plots the recurrence of the 5 most frequent nouns used by each user in the notes, translated into English. As the figure shows, each word has been used frequently by one user and seldom by other users, indicating the personal space of entities and characteristics of the conversations in the mental health domain since the topic of these conversations, i.e. the life events and situations, varies from one patient to the other.

3.2 Generation of Personal Stimuli

We extracted one sentence from each of the 92 selected narratives using an out-of-the-shelf extractive summarizer[4], and under the supervision of the psychotherapists, designed 5 templates to convert each summary and its assigned emotion or automatically detected sentiment into a coherent stimulus consisting of a common ground and a follow-up question. For each 18 one-line narrative summaries [*Summary*] with an assigned emotion [*Emotion*] by the user, two templates are defined as;

> *In the notes you left previously, I read [Summary]. You told me you felt [Emotion] for that. Do you still feel [Emotion]?*
>
> *I remember you told me that you felt [Emotion] because of [Summary]. How do you feel now?*

while, for the 61 one-line narrative summaries with automatically determined polarity [*Sentiment*], two templates are defined as;

[3] https://www.opener-project.eu/

[4] sumy Automatic text summarizer, https://pypi.org/project/sumy/

Previously, you had a [Sentiment] feeling about what I read in your note [Summary]. How do you feel about it now?

I remember you had a [Sentiment] feeling about what I read in your note [Summary]. Do you have any new thoughts or considerations about it now?

and, for the 13 one-line narrative summaries without any assigned emotion or determined polarity, one template is defined as;

I read in your note about [Summary]. Do you want to tell me more about it now?

Using this methodology, we obtained 171 stimuli from the 92 selected narratives, of which 150 stimuli are used as the grounding and conversation context for follow-up dialogue generation while 21 stimuli (approximately equal to 10% of the set) are selected by stratified sampling, as a reserved subset. Table 1 shows the statistics regarding the distribution of the stimuli type used for the dialogue generation process.

3.3 Generation of Dialogue Follow-Ups

Two dialogue writer groups were recruited for the dialogue generation. The first group included 4 psychotherapists experienced in ABC therapy technique, and the second group included 4 non-expert writers. Each writer was presented with a detailed guideline including the task description as well as several examples of correct and incorrect annotation outcomes. For each provided stimulus, the writers were asked to firstly review and validate the stimulus for possible "Grammatical Error" or "Inter-sentence Incoherence" and in case of an invalid stimulus, to apply necessary modifications to correct it. Following the validation, the writers were asked to write a short dialogue follow-up based on the stimulus, assuming that the stimulus was asked by a Personal Healthcare Agent (PHA) to a user about his/her previous narrative.

The writers were asked to respect three mandatory requirements while generating the dialogues as 1) The conversation must be based on and consistent with the stimulus; 2) The flow of the conversation must be such that the user elaborates about the event introduced in the stimulus and provides more information about the event and its objects (person, location etc.) or his/her emotion to the PHA; and 3) The conversation must contain a closure turn by the PHA. The closure turn is an important part of the generated dialogue because these sentences play the role of the acknowledgment and grounding of the dialogue between the user and the PHA, and at the same time may increase the user willingness to use the PHA. The number of turns for the dialogues was not fixed. However, the dialogue writers were suggested to write 4 dialogue turns for each stimulus, resembling 2 turns for the user and 2 turns for the PHA (excluding the stimulus) with the last turn as the closure by the PHA. Furthermore, in order to minimize cognitive workload, the writers were suggested to distribute the work by taking a break after each 10 stimuli.

Stimulus Type	Category	Count	Total Count
with Emotion	Fear	2	32
	Happiness	9	
	Sadness	10	
	Anger	7	
	Disgust	2	
	Surprise	2	
with Valence	Positive	57	107
	Negative	50	
Neutral	-	-	11

Table 1: The distribution of the stimuli used for follow-up dialogue collection, obtained by the automatic aggregation of extracted one-line summaries, the templates and the assigned emotion or automatically detected sentiment valence.

Initially, 10 stimuli were selected by stratified sampling as the Qualification Batch and were provided to all the writers for the purpose of training and resolving possible misunderstandings. The outcome of the Qualification Batch was then manually controlled and few adjustments were made with 2 of the writers. Afterwards, the rest of the stimuli were distributed such that 30% of the stimuli are annotated by all 8 writers and the rest of the stimuli are annotated by two psychotherapists and two non-expert writers.

4 Evaluation

Using the introduced elicitation methodology, we collected a corpus of follow-up conversations from the two writer groups[5]. We then performed an analysis on the obtained conversations to evaluate the

[5]We are currently applying for further funds to anonymize the corpus and publish a version of the corpus that respects patients' privacy and deontological requirements.

	Non-Experts	Therapists
# Dialogues	400	400
# Turns	1714	1494
# Unique Tokens	3146	4251
Avg. Turns per Dialogue	4.2	3.7

Table 2: The statistics of the collected corpus of follow-up dialogues using the proposed elicitation methodology per each writer group, non-experts and psychotherapists.

Dialogue Act	Non-Experts	Therapists
inform	1487	1777
answer	768	925
auto-positive	591	333
question	396	452
request	217	194
suggest	162	167
offer	117	26
confirm	65	36
disconfirm	56	63
address-suggest	40	17
address-request	2	9
other	77	11

Table 3: The distribution of the Dialogue Acts in the generated follow-up conversations by each writer group using ISO standard DA tagging in Italian (Roccabruna et al., 2020). Less frequent DAs to the task as accept-apology, apology, promise, accept-offer, and Feedback dimension DAs auto-negative, allo-negative and allo-positive are presented as "other" in the Table (Bunt et al., 2010).

elicitation methodology and to investigate the impact of domain expertise on the collected dialogues by comparing the performances of psychotherapists and non-expert writers.

4.1 Validation of the Generated Stimuli

In the first subtask, while 34.2% of the provided stimuli to the non-expert writers were labeled as invalid, this percentage by the psychotherapist group was 44.5%. Furthermore, the inter-annotator agreement measured by Fleiss κ coefficient (Fleiss, 1971) was higher in the latter group (0.26) as opposed to the non-expert group (0.06). This discrepancy in the validation subtask suggests that the assessment of the stimuli by each writer is affected by their level of competence in the domain and a more precise assessment of the stimuli as an effect of domain expertise. Therefore, domain expertise seems to be an important requirement for the quality of validation annotation in the mental health domain. Nevertheless, by representing each writer group by their consensus vote over the subset of stimuli for which we have a consensus decision, the inter-group agreement over this subset of 27 stimuli was 0.6639, measured by Cohen's κ coefficient (Cohen, 1960), suggesting that even though domain knowledge and expertise results in a fine-grained assessment, it is still feasible to obtain a course-grained validation over the generated stimuli with a group of non-expert writers with appropriate guidelines.

While the expert group labeled 60% of the invalid stimuli due to "Inter-sentence Incoherence" with respect to the automatic generation and combination of the stimuli elements (the summary, the sentiment, and the template), "Grammatical Error" was the assigned error in most of the stimuli labeled as invalid, 69%, by the non-expert group. Regarding the corrections applied to the invalid stimuli, modifications were mostly about the automatically extracted summary and detected polarity. The modifications on the summary sentence included refactoring the structure, re-positioning sections of the summary or restoring the punctuation. As for the modifications on the detected sentiment, while the modifications done by the non-expert writers were about changing negative and positive polarity with one another, the experts tended to be more conservative in expressing a sentiment for the stimuli as they mostly changed the stimuli with detected sentiment to neutral ones without any polarity.

In less than 10% of the cases the writers, mostly the psychotherapists, modified the template and specifically the follow-up question. In these cases, the questions were changed to a more summary-specific ones such as *"...What was the distorted thought that came to your mind?"*.

4.2 Analysis of the Dialogue Data Collection

As the result of elicitation process, we collected a dataset of follow-up dialogues in the mental health domain, presented in Table 2, consisting of 800 dialogues written by both groups. The number of turns and the number of unique tokens for each group indicate that the experts tended to write shorter conversations while they used a wider range of vocabulary in writing the conversations compared to the non-expert group. Regarding the length of the generated dialogues, in 627 conversations the writ-

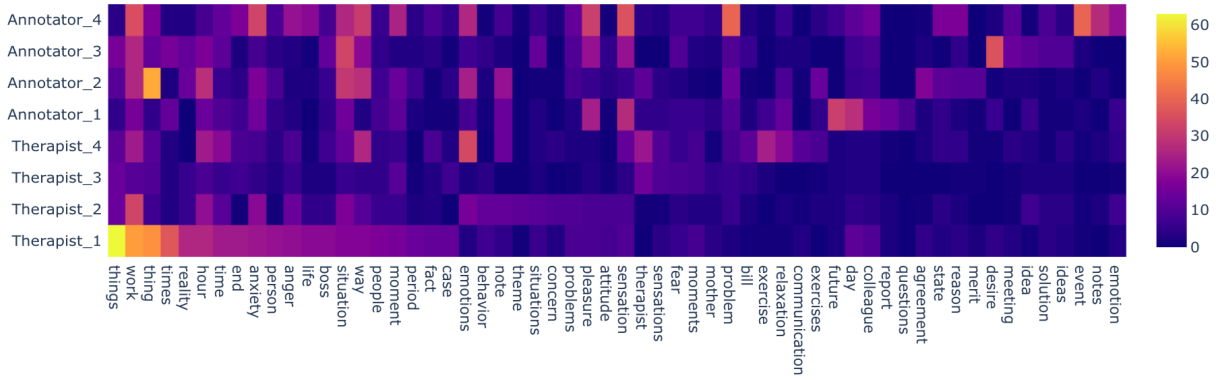

Figure 4: The heat-map of frequent nouns used by the dialogue writers in the generated conversations (English translations). The x-axis represents the nouns extracted by merging the lists of 20 most frequent nouns used per each writer. The y-axis and z-axis represent the writers and the noun frequency per each writer respectively.

ers respected the suggestion of writing 4 turns per dialogue, with exceptions of 90 dialogues written in two turns where the user replies to the stimulus and the PHA ends the conversation with a closure turn, and 83 dialogues where the user and the PHA discuss further about the event and the user's thoughts before ending the conversation.

4.2.1 Linguistic Analysis

In order to gain insights about the differences in the dialogues written be each group, we looked into the vocabulary of the nouns and entities used by each writer. Figure 4 shows the frequency heat-map of the 20 most frequent nouns used by each writer in generated dialogues, translated into English. The results indicate that the language and vocabulary used in the expert group is specific for each therapist and varies from one expert to the other, while non-expert writers have a more combined vocabulary with less inter-annotator novelty in lexicon, suggesting that the domain expertise has an influence on language and the use of vocabulary in generating conversations for the mental health domain.

Furthermore, we developed a Dialogue Act tagger to compare the conversations by their set of Dialogue Acts (DA). For this purpose, we annotated 370 of the collected dialogue follow-ups (1514 turns, approximately equal to 45% of the dataset) with the ISO standard DA tagging in Italian (Roccabruna et al., 2020) and trained an encoder–decoder model (Zhao and Kawahara, 2019) to segment each turn to its functional units and label them by their DAs. The results, presented in Table 3, show that despite the similarity in the use of the top 6 frequent DAs (inform, answer, auto-positive, question, request and suggest), there is a diversity in the type and the frequency of the DAs used by non-expert group (such as offer, address-suggest and other less relevant DAs to the domain) with respect to the professionals, suggesting that the professionals hold a more structured conversation with respect to the other group.

4.2.2 Response-Selection Baseline

We investigated the appropriateness of the collected dialogue corpus for developing conversational agents in the mental health domain by training a TF-IDF response-selection baseline model. The model was trained on 90% of the collected conversations with a similar training setting to Lowe et al. (2015), and evaluated on the remaining 10% of the data as test set using *Recall@k* family of metrics, presented in Table 4. The model was then integrated in the application introduced in subsection 3.1 to select the correct PHA response for each user turn. 10 test users were recruited to interact with our application and write narratives about their life events by answering the ABC questions for 50 days. Each narrative was then automatically converted to a personal dialogue stimuli after one day, using the introduced methodology in subsection 3.2, to initiate a follow-up dialogue with the test user for two exchanges (4 turns) with natural language responses from the users and retrieved responses from the system. Regarding the evaluation of the dialogues, we asked the test users to assess the appropriateness and coherence of each system turn (including the stimulus) during the conversation with thumbs-up (appropriate) or thumbs-down (inappropriate) for each turn, and to evaluate the quality of the conversation as-a-whole by voting

	TF-IDF
1 in 2 R@1	0.49
1 in 10 R@1	0.21
1 in 10 R@2	0.36
1 in 10 R@5	0.55
1 in 50 R@1	0.14
1 in 50 R@2	0.18
1 in 50 R@5	0.26

Table 4: The performance of the response-selection baseline on the collected dialogue follow-ups for different recall metrics.

	Count
# Dialogues	217
# 5-star	130 (60%)
# 4-star	26 (12%)
# 3-star	41 (19%)
# 2-star	8 (3%)
# 1-star	12 (6%)
# PHA Turns	651
# Thumps-Up	594 (91%)
# Thumps-Down	57 (9%)

Table 5: The results of human evaluation of the response-selection model in follow-up dialogues. The users rated each response on a binary scale (Thumbs-Up and Thumbs-Down) as well as the whole dialogue with 1-5 star score.

from 1-star (very bad) to 5-stars (very good) for each dialogue.

The results of human evaluation on the baseline dialogue model, shown in Table 5, indicate that 91% of the system turns were considered appropriate and coherent by the test users, resulting in more than 70% of the dialogues with acceptable quality, thus suggesting the usefulness and suitability of the generated dialogues using the proposed methodology for developing PHAs in the mental health domain.

5 Conclusions

In this work, we address the need for suitable dialogue corpora to train Personal Healthcare Agents in the mental health domain. We present an elicitation methodology for dialogues in the mental health domain grounded in personal recollections. Using the proposed methodology, we collected a dataset of follow-up dialogues that psychotherapists would hold with the patients to review the personal events and emotions during a CBT intervention.

Through an analysis of the collected resource following our proposed methodology, it emerged that the task of validating responses and generating dialogues in the mental healthcare domain can be performed both by using psychotherapists and non-expert dialogue writers. Therefore, it suggests the possibility of training a larger number of non-expert dialogue writers using appropriate guidelines to obtain a valid dataset with less cost while ensuring consistency in the results.

Furthermore, we investigated the appropriateness of the collected corpus for developing conversational agents in the mental health domain. We reported automatic and human evaluation of a corpus-based response-selection baseline. We found that the test users who interacted with the model over a long-term period (50 days) considered on average 91% of system turns as appropriate and coherent, resulting into 72% of dialogues with acceptable quality.

We believe the proposed methodology can be used to tackle the problem of resource scarcity in the mental health domain. In particular, our methodology can be used to obtain corpora of dialogues grounded in personal recollections for developing dialogue models in the mental health domain.

Acknowledgements

The research leading to these results has received funding from the European Union – H2020 Programme under grant agreement 826266: COADAPT.

References

Alaa A Abd-alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M Bewick, Peter Gardner, and Mowafa Househ. 2019. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, 132:103978.

Alaa A Abd-Alrazaq, Mohannad Alajlani, Nashva Ali, Kerstin Denecke, Bridgette M Bewick, and Mowafa Househ. 2021. Perceptions and opinions of patients about mental health chatbots: Scoping review. *Journal of Medical Internet Research*, 23(1):e17828.

Mohammad Rafayet Ali, Seyedeh Zahra Razavi, Raina Langevin, Abdullah Al Mamun, Benjamin Kane, Reza Rawassizadeh, Lenhart K. Schubert, and Ehsan Hoque. 2020. A virtual conversational agent for teens with autism spectrum disorder: Experimental results and design lessons. In *Proceedings of the 20th ACM International Conference on Intelligent*

Virtual Agents. Association for Computing Machinery.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an iso standard for dialogue act annotation. *Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Paul Ekman. 1992. Are there basic emotions? *Psychological Review*, 99(3):550–553.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e19.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Nishitha Guntakandla and Rodney Nielsen. 2018. Annotating reflections for health behavior change therapy. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Aiswarya Baiju, Bing Liu, Ben Gerber, Lisa Sharp, Nadia Nabulsi, and Mary Smart. 2020. Human-human health coaching via text messages: Corpus, annotation, and analysis. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 246–256.

Ben Krause, Marco Damonte, Mihai Dobre, Daniel Duma, Joachim Fainberg, Federico Fancellu, Emmanuel Kahembwe, Jianpeng Cheng, and Bonnie Webber. 2017. Edina: Building an open domain socialbot with self-dialogues. *1st Proceedings of Alexa Prize (Alexa Prize 2017)*.

Xiujun Li, Zachary C Lipton, Bhuwan Dhingra, Lihong Li, Jianfeng Gao, and Yun-Nung Chen. 2016. A user simulator for task-completion dialogues. *arXiv preprint arXiv:1612.05688*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294. Association for Computational Linguistics.

Horea-Radu Oltean, Philip Hyland, Frédérique Vallières, and Daniel Ovidiu David. 2017. An empirical assessment of rebt models of psychopathology and psychological health in the prediction of anxiety and depression symptoms. *Behavioural and cognitive psychotherapy*, 45(6):600–615.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2016. Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51.

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. Association for Computational Linguistics.

Giuseppe Riccardi. 2014. Towards healthcare personal agents. In *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges*, pages 53–56.

Gabriel Roccabruna, Alessandra Cervone, and Giuseppe Riccardi. 2020. Multifunctional iso standard dialogue act tagging in italian. *Seventh Italian Conference on Computational Linguistics (CLiC-it)*.

Diego Sarracino, Giancarlo Dimaggio, Rawezh Ibrahim, Raffaele Popolo, Sandra Sassaroli, and Giovanni M Ruggiero. 2017. When rebt goes difficult: applying abc-def to personality disorders. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, 35(3):278–295.

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. Association for Computational Linguistics.

Tianyu Zhao and Tatsuya Kawahara. 2019. Joint dialog act segmentation and recognition in human conversations using attention to dialog context. *Computer Speech & Language*, 57:108–127.