Università degli Studi di Trento

Department of Information Engineering and Computer Science

IECS Doctoral School

Response Generation in Longitudinal Dialogues

Advisor
Prof. Giuseppe Riccardi

Ph.D. Candidate
Seyed Mahed Mousavi

Review Committee
Dr. Dilek Hakkani-Tür,
Prof. Frédéric Béchet,
Prof. Giovanni Iacca

Trento, May 2023

Abstract

LONGITUDINAL DIALOGUES (LD) are the most challenging type of conversations for human-machine dialogue systems. LDs include the recollections of events, personal thoughts, and emotions specific to each individual in a sparse sequence of dialogue sessions. Dialogue systems designed for LDs should uniquely interact with the users over multiple sessions and long periods of time (e.g. weeks). Over an extended period of time, the machine should learn about the users' life-events and participants from the responses shared during each dialogue session, and create a personal user model. The acquired user model must consider individuals' states, profiles, and experiences that vary among users and dialogue sessions.

The acquisition of a dialogue corpus is the first key step in the process of training a dialogue model. There has been limited research on the problem of collecting personal conversations from users over a long period of time. Corpora acquisitions have been designed either for open-domain information retrieval or slot-filling tasks with stereotypical user models "*averaged*" among users. In contrast, the level of personalization in LDs is beyond a set of personal preferences and can not be learned from a limited set of persona statements.

Advancement in human evaluation is another required step to make progress in dialogue system research. Current automatic evaluation measures are poor surrogates, at best. There are no agreed-upon human evaluation protocols and it is difficult to develop them. As a result, researchers either perform non-replicable, non-transparent, and inconsistent procedures or, worse, limit themselves to automated metrics.

In this thesis, we study the design and training of dialogue models for LDs. Our first contribution is a methodology for data collection and elicitation of multi-session personal dialogues. Using the proposed methodology, we collect a dialogue corpus of human-machine LDs, followed by a case study in the mental health domain.

In the second contribution, we propose an unsupervised approach to automatically parse the users' responses at each interaction and construct the graph of users' personal space of events and participants. We extend this contribution further by studying the Information Status of the events in a personal narrative and introducing a novel chal-

lenging task of identifying new events.

In our third contribution, we address the problems of non-comparability and inconsistency of human evaluation tasks in the literature, and propose to standardize the human evaluation of the response generation model. We then present a detailed protocol for the task of human evaluation of generated responses.

Last but not least, we investigate whether general-purpose Pre-trained Language Models (PLM) are appropriate for the problem of grounded response generation in LDs. We experiment with different representations of the personal knowledge extracted from previous dialogue sessions of the user, including a novel graph representation of the mentioned events and participants. We present the automatic and human evaluations of the models, the contribution of the knowledge in the response generation, and the natural language generation errors by each model.

Contents

1 Introduction **1**
- 1.1 Research Challenges . . . 3
- 1.2 Contributions . . . 4
  - 1.2.1 Longitudinal Dialogue Collection . . . 4
  - 1.2.2 Personal Space Graph . . . 5
  - 1.2.3 Human Evaluation Standard . . . 5
  - 1.2.4 Response Generation in Longitudinal Dialogues . . . 5
- 1.3 Publications . . . 7

2 Longitudinal Dialogue Collection **9**
- 2.1 Background . . . 10
- 2.2 First Dialogue Session . . . 11
- 2.3 Second Dialogue Session . . . 13
  - 2.3.1 Personal Stimuli Generation . . . 13
  - 2.3.2 Follow-Up Dialogue Elicitation . . . 15
- 2.4 Evaluation of Elicited Dialogues . . . 15
  - 2.4.1 Validated Stimuli . . . 16
  - 2.4.2 Elicited Dialogues . . . 17
- 2.5 Case Study: Personal Healthcare Agent . . . 20
  - 2.5.1 Pilot Study 1: Participatory Design . . . 20
  - 2.5.2 Pilot Study 2: Randomized Controlled Trial . . . 21
- 2.6 Conclusions . . . 21

3 Personal Knowledge Extraction **23**
- 3.1 Background . . . 23
- 3.2 Personal Space Graph . . . 24
- 3.3 Evaluation . . . 26
  - 3.3.1 Italian Corpus . . . 26
  - 3.3.2 English Corpus . . . 27

3.4 New Event Detection . . . 27
3.4.1 Definition of New Event . . . 28
3.4.2 Annotation of New Event . . . 28
3.4.3 Evaluation of Annotated Corpus . . . 30
3.4.4 Baselines for New Event Detection . . . 31
3.5 Conclusion . . . 34

4 Human Evaluation Protocol **37**
4.1 Background . . . 38
4.2 The HE Annotation Protocol . . . 38
4.2.1 Task Design . . . 39
4.2.2 Annotator Recruitment . . . 43
4.2.3 Task Execution . . . 44
4.2.4 Annotation Reporting . . . 44
4.3 Validation of the Protocol . . . 45
4.3.1 Implementation . . . 46
4.3.2 Annotation Statistics . . . 47
4.3.3 Evaluation Results . . . 47
4.4 Conclusion . . . 49

5 Response Generation in Longitudinal Dialogues **51**
5.1 Background . . . 52
5.2 Experiments . . . 53
5.2.1 Models . . . 53
5.2.2 Dataset . . . 53
5.2.3 Grounded Response Generation . . . 54
5.3 Evaluations . . . 54
5.3.1 Automatic Evaluation . . . 54
5.3.2 Human Evaluation . . . 56
5.3.3 Generation Explainability . . . 59
5.4 Conclusion . . . 60

6 Conclusions **63**
6.1 Limitations & Future Directions . . . 64

Bibliography **65**

CHAPTER 1

Introduction

Current state-of-the-art human-machine dialogue systems cannot carry out engaging and coherent multi-session conversations unique to each individual. The state-of-the-art dialogue systems are designed for assisting the user to execute a task, holding limited chit-chat conversations with shallow user engagement, or information retrieval over a finite set of topics. Such systems are designed with the assumption of stand-alone dialogue sessions between the user and the system; the sessions are disconnected and the shared information across sessions is negligible and close to none. The personalization in these systems is limited to a stereotypical user model. This user model is implicitly inferred, from conversations with many users, or is limited to a superficial list of persona statements (e.g., "He likes dogs").

Longitudinal Dialogue (LD) is one of the most challenging types of conversation for human-machine dialogue systems. LDs are multi-session interactions that encompass user-specific situations, thoughts, and emotions. Dialogue systems designed for LDs should interact with each user uniquely over multiple sessions and long periods of time (e.g. weeks). Through each session in LDs, the dialogue system must learn about the user's personal space of events and participants and social interactions, and engage the user in personal dialogues regarding their thoughts, feelings, and personal and world events.

Figure 1.1 shows an example of three types of human-machine dialogues for two sessions in two consecutive days. The dialogue types are task-based, open-domain chit-chat, and LD. The user interactions with the tasked-based dialogue system are limited to executing certain pre-defined tasks. Regardless of how successful the model is in executing such tasks, the interaction consists of either independent/disconnected exchanges such as Day 1, or a short multi-turn dialogue such as Day 2. The user model

Time	Task-Based Dialogue	Open-Domain Chit-Chat	Longitudinal Dialogue
Day 1	**User:** Set an alarm for Monday 8 a.m.	**User:** Do you like Shining?	**User:** My boss set a meeting with me tomorrow. I am a bit worried to meet her.
	Machine: I saved the alarm.	**Machine:** *The* Shining? Yes I like most of Kubrick's movies.	**Machine:** Why? Did she say what the meeting is going to be about?
	User: Remind me to call Mike at 5 p.m.	**User:** Cool! I am going to watch the shining for the first time.	**User:** I think she wants to fire me. She has been searching for an other cashier recently.
	Machine: Reminder Saved.	**Machine:** I think you are going to enjoy it.	**Machine:** Sometime uncertainty makes us anxious. Maybe there is another motivation. No need to worry now.
Day 2	**Machine:** How may I help you?	**Machine:** What's the most interesting thing you've heard recently?	**Machine:** I remember you had a meeting with your boss today. How did it go?
	User: I want an expensive italian restaurant.	**User:** I learned that the Queen of England's last corgi died	**User:** Oh it went well haha! She wants me to train a new cashier. She will be my colleague.
	Machine: I found Scrigno del Duomo in the center. Should I book a table?	**Machine:** Wow! Was it a cardigan welsh corgi?	**Machine:** Good! So no bad news! Are you relaxed now?
	User: Yes, for two. Tonight at 8pm.	**User:** No it was pembroke welsh corgi!	**User:** Yeah. They are planning to expand the store. I am not fired! haha!
	Machine: Done! I booked a table for two at Scrigno del Duomo for 8 pm.	**Machine:** It was called Candy I think.	**Machine:** I am glad you feel better! Sometimes it's hard to control our emotions.

Figure 1.1: Examples of a task-based dialogue, a chat-chit, and a Longitudinal Dialogue (LD) in two different sessions. In comparison to LD, the sessions of the other two dialogues are disconnected as the machine does not construct a personal user model and the topics of these dialogues are not user-specific. On the contrary, the dialogue system in LD learns about the user in a timely manner and engages her in a personal dialogue encompassing her life events, thoughts, and emotions.

in this system is not personal as it adopts a stereotypical model -implicitly - inferred from dialogue corpora with multiple users. In the open-domain chit-chat dialogue, the dialogue does not include the execution of any explicit task, and the model engages the user in a conversation about movies and news. A common characteristic of task-based and open-domain dialogues is the fact that there is no personal information carried to the next dialogue session. The system does not update/modify the user model with each dialogue session and the level of personalization is intact from one interaction to the other (Personalization in the natural language processing and dialogue models could be added based on the voice user interface requirements and could include the exploitation of personal information such as contact directory, preferences, etc.).

In contrast, the model designed for the LD must account for three main differences compared to the other two systems; A) the contents of the LD are not about general information or knowledge matters as LDs encompass personal emotions, user and time-specific situations, and participants; B) the sessions are not disconnected dialogues and we can not model them as stand-alone interactions. In contrast, they belong to a multi-session interaction unique to the individual user, where the information shared in each interaction creates a common ground between the machine and the user. For each interaction, the system must engage the user in a dialogue respecting the common ground based on the information shared in the previous interactions, as well as the novel information in the new dialogue history; C) the machine has to extract the personal information presented in the user responses to construct and update the user model

and respond coherently. Similar to a natural interaction between human speakers, the model has to gradually become acquainted with the user throughout the dialogues and not from a superficial list of sentence-based persona descriptions.

In this work, we study the task of response generation in LDs. Response generation in LDs is subject to appropriateness and accuracy as well as personalization and engagement of the user. The level of personalization in LDs is beyond a set of personal preferences and can not be learned from a limited set of persona statements ("*I like cars*" does not necessarily imply that I like to talk about cars in my interactions). The generated response needs to respect individuals' states, profiles, and experiences that vary among users and dialogue sessions. Therefore, we can not collect a massive knowledge base of user models that can suit all individuals and scenarios. The dialogue system should learn about each user and generate a personal response that is coherent with respect to the dialogue context as well as the previous dialogue sessions.

We investigate the applicability of general-purpose Pre-trained Language Models (PLM) for grounded response generation in LDs. PLMs have achieved comparably well performance as end-to-end generative models for open-domain chit-chats [97], goal-oriented agents [79] or question answering about a finite set of topics [99]. We study whether PLMs can generate a response that is coherent with respect to the dialogue history and grounded on the personal knowledge the user has shared in previous interactions.

1.1 Research Challenges

We encountered three main research challenges throughout our studies of LDs.

Challenge 1: Data The acquisition of a dialogue corpus is a key step in the process of training a dialogue model. One of the major reasons for the limitations of state-of-the-art dialogue systems for LDs is the lack of dialogue data. There has been scarce research on the problem of collecting personal conversations with users over a long period of time. Engaging the user to elaborate on personal situations and emotions is a challenging task and designing appropriate collection/elicitation methodologies is not straightforward. The two main approaches to collecting dialogue data are a) acquiring user interaction data via user simulators and hand-designed policies [35], and b) collecting large sets of human-human conversations in different user-agnostic settings. These approaches have been used for goal-oriented agents and slot-filling tasks (e.g. reservations of restaurants) or open-domain information retrieval 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 multi-session conversations over several weeks or months.

Challenge 2: User Modeling As mentioned previously, most dialogue systems either assume a stereotypical user model "*averaged*" among all users or settle for a list of superficial persona statements. However, the situations and feelings that the LDs encompass are unique to each individual user and dialogue session. This level of personalization and user modeling can not be crowd-sourced or implicitly inferred from huge amounts of data. In these dialogues, the user responses are rich with emotions and

experiences through each exchange and turn that is unique to each user. Therefore, the model has to be able to extract this knowledge and learn about the user and derive the individual user model through/from the dialogue sessions and user responses.

Challenge 3: Evaluation Proper evaluation of dialogue models is necessary for the advancement of human-machine dialogue research. Automatic evaluation measures are poor surrogates, at best. Several studies have shown that automatic metrics can not be good candidates for evaluating a dialogue model [37, 70]. Therefore, Human Evaluation (HE) is still the necessary approach to evaluate response generation models [75]. Nonetheless, little attention has been given to the assessment of the design of HE task. Due to the lack of an agreed-upon and standard protocol, dialogue systems have been evaluated with different granularity (turn-level vs dialogue-level), different evaluation policies (single-model vs pairwise-model, candidate-ranking vs. winner-selection) and in different modalities (interactive vs static) [75]. As an outcome, countless HE tasks have been presented and conducted, resulting in non-transparent procedures, non-replicable and incomparable results, and unclear resource allocations.

1.2 Contributions

1.2.1 Longitudinal Dialogue Collection

To address the lack of proper dialogue data, we studied LDs and multi-session conversations in the mental health domain. In this domain, therapists deliver interventions over a long period of time and need to monitor or react to patients' input. During the interventions, the therapists initiate user-specific dialogues to 1) follow up with the patient and monitor the progress regarding the events mentioned in previous interventions, and 2) learn about novel life events of the narrator as well as his/her corresponding thoughts and emotions. These dialogues take place in a timely manner and encompass user-specific events, thoughts, and emotions in a complex structure. Therefore, they accurately present the complexity of dialogues a system trained for LDs should handle.

We propose a novel methodology to collect corpora of human-machine LDs [48]. Using the proposed methodology, we collected a corpus of LDs consisting of 800 two-session dialogues for each individual user. In the first dialogue session, the user recollects daily life events that she has experienced in a system-initiated conversation. For each user, the first session is then followed by a second dialogue session. In the second dialogue session, the user tends to share more details about her feelings and the possible evolution of the mentioned events, while the listener provides personal suggestions and asks questions to expand or disambiguate previously stated facts or feelings.

Case Study: Personal Healthcare Agent

As a case study in this contribution, we designed and developed a personal healthcare agent using the collected dialogue corpus. This study was in collaboration with a team of psychotherapists. The developed agent would engage the users in a dialogue for recollecting the real-life events that activated their emotional states and provide support. It would further engage each individual user in a second dialogue session to follow up

on the progress of her thoughts and emotions. The developed healthcare agent was deployed in two clinical pilot studies, including the first registered Randomized Control Trial using a dialogue system application [10, 11].

1.2.2 Personal Space Graph

To address the need for appropriate user modeling and personalization, we propose an unsupervised approach to automatically extract personal knowledge from the user's responses in previous dialogue sessions [49]. Using the definition of event based on the verb and its linguistic dependencies [5], we automatically parse the user's responses and extract the mentioned life events and their participants. We present the extracted information as a user-specific graph in terms of events as the edges of the graph, and the participants as the nodes of the graph. This graph is then used as personal knowledge for the task of response generation.

As a further contribution, we study how "new" an event is with respect to the discourse stretch. To obtain salient information from the user narrative, it is necessary to distinguish new events and the ones that have been mentioned in the current or previous dialogue sessions. We study the Information Status [58] of the events and propose a novel challenging task: the automatic identification of *new* events in a narrative [51]. We consider an event new if 1) it provides novel information to the reader with respect to the discourse (discourse-new) and 2) such information can not be inferred through commonsense. We annotated a complete dataset of personal narratives with new events at the sentence level using human annotators. We then developed several neural and non-neural baselines for the task of new event detection in both settings of candidate selection and sequence tagging. We publish the annotated dataset, annotation materials, and machine learning baseline models for the task of new event extraction for narrative understanding.

1.2.3 Human Evaluation Standard

To address the incomparability and ambiguity in HE tasks, we propose to standardize the experimental methodology for the HE of response generation models [50]. We present a protocol to the community for this task, in order to increase the comparability, replicability, and interpretability across research reports. Our proposal includes all the required steps and materials to conduct HE in a transparent and extendable way (including task design, annotator recruitment, task execution, and annotation reporting). The proposed protocol is domain-agnostic, language-independent, and open to collaborative extensions from the research community to different versions and standards.

1.2.4 Response Generation in Longitudinal Dialogues

We study the task of response generation in LDs. We investigate the applicability of general-purpose Pre-trained Language Models (PLM) for this purpose. We conversationally fine-tune two recent PLMs, GePpeTto (GPT-2) [12] and iT5 [72], as a decoder-only and an encoder-decoder architecture. To improve the quality of machine

responses, we experiment with grounded response generation. We use the user responses in the previous dialogue sessions as personal knowledge and experiment with different representations of the knowledge piece, including the graph representation of the mentioned events and participants. We evaluate the performance of the models and the impact of different knowledge representations through automatic evaluations, including explainability studies using the Integrated Gradients technique, [77], as well as HE, including the categorization of natural language errors by each model.

The contributions of this dissertation can be summarized as follows:

- **Resources**:
  - A methodology for data collection and elicitation of LDs, as well as a dataset of LDs consisting of 800 two-session human-machine dialogues.
  - An approach to automatically extract personal life events and participants from user responses and construct the graph of the user's personal space. This graph can be used as personal knowledge for user modeling and personalized response generation.
  - A novel task of new event detection in the narratives as well as the annotated version of a public corpus of personal narratives at sentence level along with its annotation methodology and evaluation.
  - Baseline benchmarks for the task of new event extraction based on discourse heuristics and deep neural networks, in two different settings of candidate selection and sequence tagging.
  - A detailed protocol for human evaluation task to be used as-is, as-a-whole, in-part, or modified and extended by the research community. The proposal includes the task design, annotators recruitment, task execution, and annotation reporting.
- **Dialogue Management & Response Generation**:
  - We investigate the suitability of the collected corpus of LDs for developing conversational agents to carry out LDs.
  - We study the task of response generation in LDs.
  - We conversationally fine-tune two PLMs with and without grounded response generation on personal knowledge. We study the performance of the models and how different representations of knowledge can affect generation quality.
  - We evaluate and compare the performance of the models using automatic evaluation, including explainability studies, and human evaluations, including studying the sub-dimensional errors made by each model.

1.3 Publications

The findings of this thesis are partially published in the following articles:

- **International Journals**
  - Danieli, M., Ciulli, T., **Mousavi**, S. M., Silvestri, G., Barbato, S., Di Natale, L., & Riccardi, G. (2022). *Assessing the Impact of Conversational Artificial Intelligence in the Treatment of Stress and Anxiety in Aging Adults: Randomized Controlled Trial*. JMIR Mental Health, 9(9), e38067.
  - Danieli, M., Ciulli, T., **Mousavi**, S. M., & Riccardi, G. (2021). *A Conversational Artificial Intelligence Agent for a Mental Health Care App: Evaluation Study of Its Participatory Design*. JMIR Formative Research, 5(12), e30053.
- **International Conferences/Workshops with Peer Review**
  - **Mousavi**, S. M., Roccabruna, G., Lorandi, M., Caldarella, S. & Riccardi, G. (2022, December). *Evaluation of Response Generation Models: Shouldn't It Be Shareable and Replicable?*. In Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), (pp. 136–147), Association for Computational Linguistics.
  - **Mousavi**, S. M., Roccabruna, G., Tammewar, A., Azzolin, S., & Riccardi, G. (2022, May). *Can Emotion Carriers Explain Automatic Sentiment Prediction? A Study on Personal Narratives*. In Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (pp. 62-70), Association for Computational Linguistics.
  - **Mousavi**, S. M., Cervone, A., Danieli, M., & Riccardi, G. (2021, June). *Would you like to tell me more? generating a corpus of psychotherapy dialogues*. In Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations (pp. 1-9), Association for Computational Linguistics.
  - **Mousavi**, S. M., Negro, R., & Riccardi, G. (2021). *An Unsupervised Approach to Extract Life-Events from Personal Narratives in the Mental Health Domain*. In Eighth Italian Conference on Computational Linguistics (CLiC-it).
- **Publications Under Review**
  - **Mousavi**, S. M., Caldarella, S., & Riccardi, G. *Response Generation in Longitudinal Dialogues: Which Knowledge Representation Helps?*
  - **Mousavi**, S. M., Tanaka, S., Yoshino, Nakamura, S. & Riccardi, G. *What's New? Identifying the Unfolding of New Events in a Narrative.* arXiv preprint arXiv:2302.07748 (2023).

CHAPTER 2

Longitudinal Dialogue Collection

Research on engaging the user in multi-session conversations over a long period of time is very scarce. Designing appropriate collection/elicitation methodologies to engage the user in the recollection of personal situations and emotions is a challenging task and not straightforward. As a result, research on multi-session dialogues resorts to crowd-sourcing datasets with superficial persona statements and pretended longitudinality [1, 90, 91]. Meanwhile, studies on LDs have been limited to inferring user's attributes such as age and gender [85], or next quick-response selection from a candidate set of "yes," "haha," "okay," "oh," and "nice" [84]. Currently, available dialogue corpora (that we will review in this chapter) are not suitable to train a dialogue system for Longitudinal Dialogues (LD). Moreover, current approaches for corpora acquisitions have been designed for open-domain chit-chat, question-answering, or slot-filling tasks with stereotypical user models.

We address the need for suitable corpora of LDs by proposing a novel methodology to collect corpora of such dialogues. We study follow-up dialogues in the mental health domain that a psychotherapist would initiate in reviewing the progress of the intervention. In this domain, the therapists deliver multi-session interventions with the individual user and follow the development of user-specific life events and emotions. During each intervention session, the therapists initiate follow-up dialogues with the users to monitor the evolution of the previously mentioned events/emotions. As previously mentioned, these dialogues present the complexity of LDs accurately, as they take place in a timely manner and encompass user-specific events, thoughts, and emotions in a complex structure.

2.1 Background

Open-domain dialogue corpora Previously published research has addressed the problem of collecting dialogue data starting from world knowledge facts or predefined persona descriptions. In this regard, Gopalakrishnan et al. [18] 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. Rashkin et al. [64] crowdsourced a dataset of conversations with implied user feelings in the context, using AMT workers, where a worker writes a personal situation associated with 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 multi-session and longitudinal dialogues.

Personal Dialogue Research on personalized response generation has focused on persona descriptions and synthetic sets of user preferences and profiles. Zhang et al. [95] collected Persona-Chat dataset of open-domain dialogues using crowd workers using Amazon Mechanical Turk (AMT) workers, where the workers were instructed to impersonate speakers with synthetic personas of 5 sentences. This dataset has been studied for personal response generation by fine-tuning PLMs [29, 87], by learning the users' persona from the dialogues samples rather than the persona descriptions [41], or investigating different representations of persona statements [24]. While the mentioned work focused on personalization in open-domain dialogues, Joshi et al. [27] generated profiles consisting of gender, age, and food preference permutations for the user side in restaurant booking dialogues, which was used in another work [74] to generate personalized responses in a task-based dialogue.

Multi-session Dialogue Studies on multi-session dialogues have been limited to simulated longitudinality and superficial persona. Xu et al. [90] extended the Persona-Chat dataset to a multi-session chat dataset with 4 to 5 sessions, by instructing crowdworkers to impersonate the role of returning dialogue partners in the first session (extracted from the Persona-Chat dataset) after a random amount of time. The workers were explicitly asked not to discuss any personal and real-life matters but play the role defined by the persona statements. This approach was further used by Bae et. al. [1] to extend an existing dataset of persona chats in Korean to multi-session dialogues. Xu et al. [91] proposed a framework for persona memory in multi-session dialogues and collected a dataset of persona chats in Chinese via crowd workers.

Therapy dialogue corpora The research in this domain is very recent and resources are scarce. "Counseling and Psychotherapy Transcripts" published by Alexander Street Press[1] is a dataset of 4000 therapy session transcriptions on various topics, used as a resource for therapists-in-training. Pérez-Rosas et al. [57] 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 [19] conducted a data collection process of therapeutic dialogues in Wizard of Oz manner where the therapists impersonated a healthcare agent. The authors

[1] https://alexanderstreet.com/

recorded 324 sessions of therapeutic dialogues which were then manually transcribed. Furthermore, in the physical health coaching domain, Gupta et al. [20] collected a dataset of conversations where the expert impersonates a healthcare agent that engages the users into a healthier lifestyle. For this purpose, a certified health coach interacted with 28 patients using a messaging application.

2.2 First Dialogue Session

The type of dialogue that we aim at obtaining is different from what has been reported in the literature. We present an elicitation methodology to generate a dataset of LDs, encompassing real-life events and emotions which vary among users and dialogue sessions. The LDs collected in this work consist of two dialogue sessions for each individual user. For the first dialogue session, we collect a dataset of personal human-machine conversations about user-specific life events and participants. A group of 20 Italian native speakers who were receiving Cognitive Behavioral Therapy (CBT) was asked to interact with a dialogue system and write notes about the daily events that activated their emotional state[2].

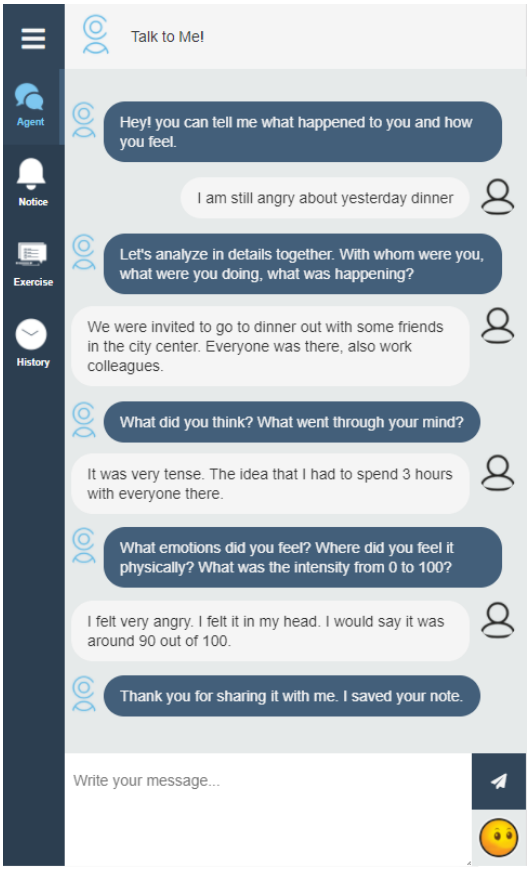

Figure 2.1: The user interface of the mobile application designed for collecting the first dialogue sessions (English translations). The patients were asked to interact with the dialogue agent and answer the ABC questions designed by psychotherapists.

CBT is a psychotherapy technique based on the theory 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 [54]. A technique commonly used in CBT treatment is the ABC (Antecedent, Belief, Consequences). 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

[2]This study has been approved by the Institutional Review Board of the University of Trento.

2.2. First Dialogue Session

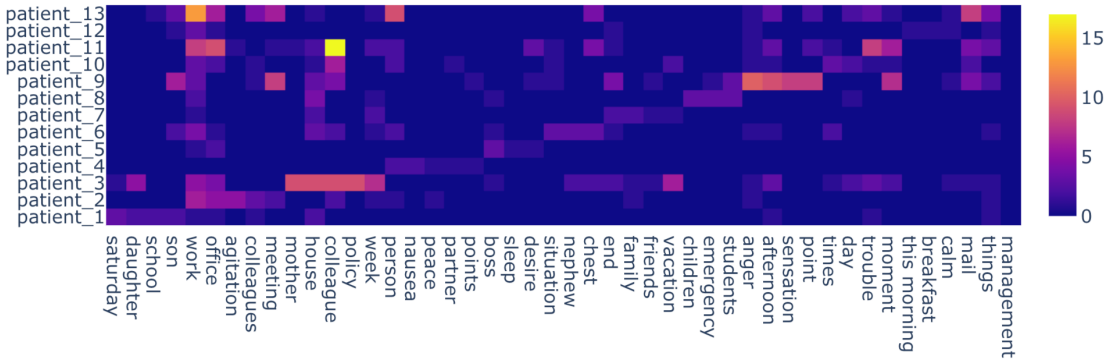

Figure 2.2: The heat-map of frequent nouns used by the users in the first dialogue sessions (English translations). The x-axis represents the nouns extracted from the 5-most frequent list used by each user while the y-axis and z-axis represent the users and the noun frequency, respectively.

functional thoughts [71].

We recruited 20 users who would meet with their human psychotherapists one session a week and asked them to write notes about the daily life events that activated their emotional state between one session and the following one. For this purpose, we designed a rule-based conversational agent as a mobile application that the users could interact with for a period of three months, to answer the four questions designed by the psychotherapists for the ABC technique, and assign an emotion to the note. The machine turns were randomly selected from a set of 3 templates for each question A,B, and C (the templates were different lexicalizations of the same question designed by the therapists). 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) [15], and two other complex emotional states (Embarrassment and Shame) that were considered relevant for this setting. Figure 2.1 shows the user interface of the application designed for this purpose.

By the end of this step, 224 ABC dialogues were obtained from 20 users of which 92 dialogues (written by 13 different subjects) were complete, i.e. the users have answered all the questions completely. Lexical analysis of the complete dialogues demonstrates that the language and vocabulary used in the user responses are user-specific. Figure 2.2 plots the recurrence of the 5 most frequent nouns used by each user in their responses, translated into English. As the figure shows, each word has been used frequently by one user and seldom by other users. This result indicates the level of personalization is the users' space of entities and characteristics in the conversations. Therefore, in contrast to task-based or open-domain dialogues, the topic of these conversations, i.e. the life events and situations, varies from one user to the other. Nevertheless, nouns such as *office* and *work* are used frequently by all users, suggesting that they can be the common reasons for emotion activation among the users.

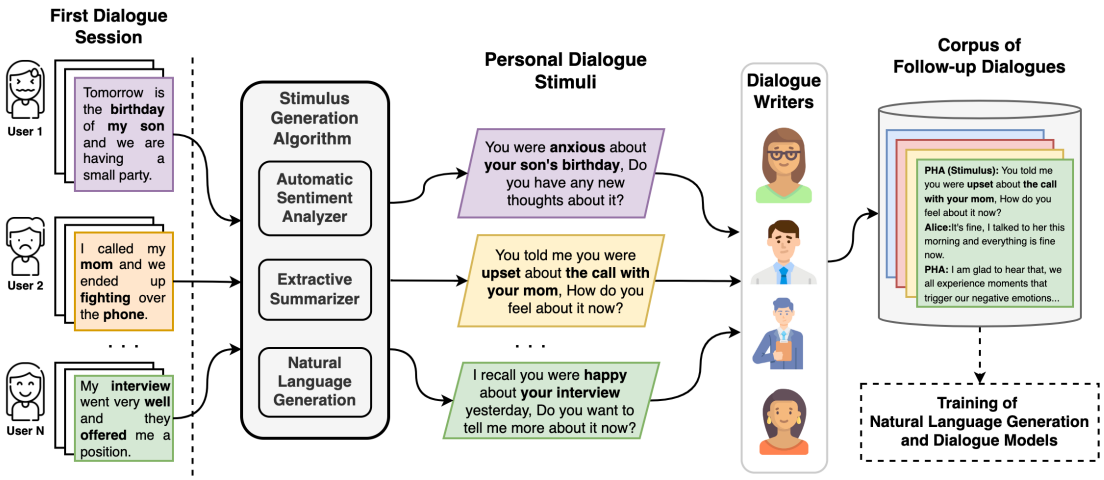

Figure 2.3: The workflow for the elicitation of follow-up dialogues, starting from the user responses in the first dialogue session. The stimulus generation algorithm creates a personal dialogue stimulus as a seed for dialogue writers. The writers use the textual stimulus and principled guidelines to generate the follow-up dialogues.

2.3 Second Dialogue Session

We used the collected first-session dialogues as the context to elicit dialogue follow-ups for each user. We generated personal dialogue stimuli for follow-up conversations grounded in user responses in the first dialogue session. The stimuli were presented to domain experts (therapists) and non-expert dialogue writers and they were asked to generate a human-machine dialogue, as a second dialogue session, by impersonating themselves as both sides of the conversation. Figure 2.3 presents the proposed workflow for the acquisition of follow-up dialogues as the second dialogue sessions.

2.3.1 Personal Stimuli Generation

The stimuli consist of two parts. The first part of the stimulus, the common-ground statement, contains the summary of what the user shared in the first session and the associated emotions, while the second part is a follow-up question aimed at reviewing the user's life events.

The user responses in the first-session dialogue are related to each other and present details about the same event. Therefore, we concatenated the user responses in each dialogue under the therapists' supervision to convert them into personal narratives of one piece. We extracted one sentence from each of the 92 selected narratives using an extractive summarizer for the Italian language based on the Latent Semantic Analysis technique [76] to obtain the most representative sentence.

Out of the 92 complete dialogues, 18 dialogues were assigned an emotion by the user, and 74 dialogues were not labeled with any emotions. We used a lexicon-based sentiment analyzer to classify the 74 narratives without any expressed emotions by polarity. The model classified 61 narratives as either negative or positive and 13 of

2.3. Second Dialogue Session

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 2.1: The distribution of the stimuli used for follow-up dialogue elicitation, obtained by the automatic aggregation of extracted one-line summaries, the templates and the assigned emotions or automatically detected sentiment polarities.

them as neutral.

Under the supervision of the psychotherapists, we 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 the narratives with an assigned [*Emotion*] by the user, two templates were defined:

> *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 narratives with automatically determined polarity [*Sentiment*], two templates were defined;

> *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 narrative summaries without any assigned emotion or determined polarity, one template was defined;

> *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. We then reserved 21 stimuli (approximately equal to 10% of the set) selected by stratified sampling, as backup subset. The remaining 150 stimuli were used as the stimulus and conversation context for follow-up dialogue elicitation. Table 2.1 shows the statistics regarding the distribution of the stimuli used for the dialogue elicitation process.

2.3.2 Follow-Up Dialogue Elicitation

We recruited two dialogue writer groups for the elicitation of second-session dialogues. The first group included 4 psychotherapists experienced in the 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 first 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 follow-up dialogue based on the stimulus, assuming that the stimulus was asked by a dialogue agent to the user regarding her previous interaction.

The writers were asked to respect three mandatory requirements while generating the dialogues:

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 on the event introduced in the stimulus and provides more details about the event (person, location etc.) or her emotion;
3. The conversation must contain a closure turn by the agent.

The closure turn is an important part of the dialogue because these sentences play the role of the acknowledgment and grounding of the dialogue between the user and the agent, and at the same time may increase the user's willingness to interact with the agent. 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 agent (excluding the stimulus), with the last turn as the closure by the agent. Furthermore, in order to minimize cognitive workload, the writers were suggested to distribute the work by taking a break after every 10 stimuli.

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 a few adjustments were made with 2 of the writers. Afterward, 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.

2.4 Evaluation of Elicited Dialogues

Using the introduced elicitation methodology, we collected a corpus of follow-up conversations from the two writer groups, presented in Table 2.2. We evaluated the elicitation methodology and investigate the impact of domain expertise on the collected dialogues by comparing the performances of psychotherapists and non-expert writers.

The number of turns and the dictionary size for each group indicate that the experts tend to write shorter conversations while they used a wider range of vocabulary in the

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

Table 2.2: The statistics of the collected corpus of follow-up dialogues using the proposed elicitation methodology per each writer group, non-experts and psychotherapists.

conversations compared to the non-expert group. Regarding the length of the generated dialogues, in 627 conversations the writers respected the suggestion of writing 4 turns per dialogue. Regarding the other 173 dialogues, for 90 dialogues two turns are written where the user replies to the stimulus and the dialogue agent ends the conversation with a closure turn, while, 83 dialogues consist of more than 4 turns as the user and the agent discuss the event and the user's thoughts further before ending the conversation.

2.4.1 Validated Stimuli

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%. Besides, the inter-annotator agreement measured by Fleiss κ coefficient [16] was higher in the expert group (0.26) compared 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, i.e. domain expertise leads to a more precise assessment of the stimuli. Nevertheless, by representing each writer group by their consensus vote over the subset of stimuli annotated by all writers, the inter-group agreement over this subset of 27 stimuli is 0.66, measured by Cohen's κ coefficient [8]. This result suggests that even though domain knowledge and expertise result in a fine-grained assessment, it is still feasible to obtain a course-grained validation of the generated stimuli with a group of non-expert writers using appropriate guidelines.

Regarding the type of errors annotated by each group, 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). Meanwhile, 69% of the stimuli labeled as invalid by the non-expert group were due to "Grammatical Error". Regarding the corrections applied to the invalid stimuli, modifications were mostly about the automatically extracted summary and detected polarity. The modifications to 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 mostly about changing negative and positive polarity with one another, the experts tended to be more conservative in expressing sentiment for the stimuli as they mostly changed the stimuli with detected sentiment to neutral. 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

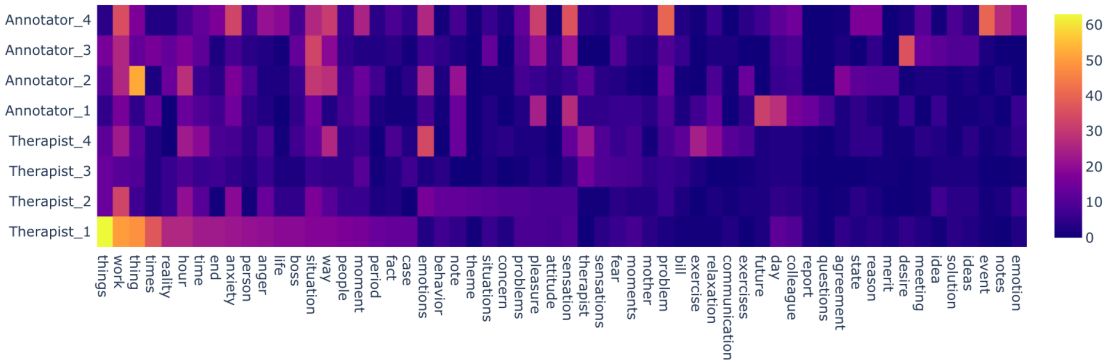

Figure 2.4: The heat-map of frequent nouns used by the dialogue writers in the generated conversations (English translations). The x-axis represents the nouns extracted by merging the lists of 20 most frequent nouns used per each writer. The y-axis and z-axis represent the writers and the noun frequency per each writer respectively.

more summary-specific ones such as *"...What was the distorted thought that came to your mind?"*.

2.4.2 Elicited Dialogues

In order to gain insights into the differences in the dialogues written by each group, we looked into the vocabulary of the nouns and entities. Figure 2.4 shows the frequency heat-map of the 20 most frequent nouns used by each writer in elicited dialogues, translated into English. The results indicate that the language and vocabulary used in the expert group are specific for each therapist and vary from one expert to the other. Meanwhile, non-expert writers have a more combined vocabulary with less inter-annotator novelty in the lexicon. This result suggests that domain expertise has an influence on language and the use of vocabulary in generating conversations.

Dialogue Act Tagger

We developed a Dialogue Act tagger to compare the elicited dialogues by their set of Dialogue Acts (DA). For this purpose, we annotated 370 of the collected dialogues (1514 turns, approximately equal to 45% of the dataset) with the ISO standard DA tagging in Italian [68] and trained an encoder–decoder model [98] to jointly perform functional unit segmentation and dialogue act tagging. The results, presented in Table 2.3, 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.

Response Selection Baselines

We investigated the appropriateness of the elicited follow-up dialogues for developing human-machine dialogue systems. For this purpose, we developed four response-

2.4. Evaluation of Elicited Dialogues

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 2.3: The distribution of the Dialogue Acts (DA) in the elicited follow-up conversations by each writer group using ISO standard DA tagging in Italian [68]. Less frequent DAs such as accept-apology, apology, promise, accept-offer, and Feedback dimension DAs (auto-negative, allo-negative, and allo-positive) are presented as "other".

selection baselines, as two Information Retrieval models and two Deep Neural Network models. We chose the selection setting compared to generation since 1) in this setting the system outputs the exact turns, suggestions, questions, and closures elicited from the dialogue writers, thus we directly evaluate the data and not the generation ability of the model; 2) compared to response generation, the system is always limited to a predefined list of response candidates which may not contain any appropriate response given a specific history. However, it provides a higher level of control on the model outputs, which is a mandate in the mental health domain.

The two Information Retrieval models are :

- **TF-IDF** Term Frequency-Inverse Document Frequency method, which is a common statistical method that measures how relevant a term in a document is. The intuition behind is that a term is of high importance in a document if it is seen often in that document and not so frequently in other documents in the data set.
- **BM25** Best Matching 25 [67], an Information Retrieval algorithm similar to TF-IDF which further penalizes the term frequency score based on the length of the document.

while the two Deep Neural Network models are:

- **SNN** Siamese Neural Network [40], which consists of twin networks with tied weights and computes the similarity score between the two input sequences. For each dialogue history *h*, and response candidate *r*, our model embeds the conversation history and the response candidate. It computes the similarity of the two as $\sigma = (h^T M r)$, where *M* is a matrix of parameters learned by the model during the training. The similarity score is forwarded to the sigmoid activation function to be converted to a value from 0 to 1, representing the probability that the two are

	TFIDF	BM25	SNN	SMN
1 in 2:				
R@1	0.49	**0.62**	0.54 (LSTM)	0.56
1 in 10:				
R@1	0.21	**0.24**	0.09 (GRU)	0.13
R@2	0.36	**0.37**	0.20 (GRU)	0.26
R@5	0.55	**0.63**	0.52 (LSTM)	0.60
1 in 50:				
R@1	0.14	**0.17**	0.02(LSTM)	0.03
R@2	0.18	**0.21**	0.04 (GRU)	0.06
R@5	0.26	**0.31**	0.10 (LSTM)	0.18

Table 2.4: Automatic evaluation of the response selection baselines using the elicited corpus by Recall@k metric.

	TFIDF	BM25
#Dialogues	217	52
#5-star	130 (60%)	15 (29%)
#4-star	26 (12%)	17 (33%)
#3-star	41 (19%)	8 (15%)
#2-star	8 (3%)	7 (13%)
#1-star	12 (6%)	5 (10%)
#Turns	651	107
#Th. Up	594 (91%)	72 (67%)
#Th. Down	57 (9%)	35 (33%)

Table 2.5: Human evaluation of the two outperforming models in follow-up dialogues. The users rated each response on a binary scale as well as the whole dialogue with scores from 1 to 5.

a valid pair based on their similarity as *p(y=1|h,r)*. Regarding the hidden units, we experimented with Long-Short Term Memory (LSTM) units as well as Gated Recurrent Units (GRU).

- **SMN** Sequential Matching Network [88] as a context-based matching model. The input sequences are encoded using a Recurrent Neural Network. The obtained representations are passed to a convolutional layer followed by a pooling layer. Each sequence is then presented as a vector consisting of the features extracted in the previous step. Finally, the hidden states of GRU are used to compute the final matching score for the history and the response candidate.

The models were trained on 90% of the collected conversations and evaluated on the remaining 10% of the data as test set using *Recall@k* family of metrics. The results of the automatic evaluation of the models, presented in Table 2.4, indicate that BM25 outperforms the other alternatives in all settings (the parameters of BM25 model were optimized as b=0.75 and k1=1.49.).

For the next analysis, we integrated TF-IDF and BM25 models into the application used to collect the first dialogue sessions. We recruited 10 test users to interact with our dialogue agent and hold dialogues about their life events by answering the ABC questions for 50 days. Each dialogue was then automatically converted to a personal stimulus after one day, using the previously introduced methodology. The system would use the stimulus to prompt the user and initiate a follow-up dialogue for two exchanges (4 turns) with natural language responses from the users and retrieved responses from the agent. We asked the test users to assess the appropriateness and coherence of each machine turn in follow-up dialogues (including the stimulus) with thumbs-up (appropriate) or thumbs-down (inappropriate), and to evaluate the quality of the conversation as-a-whole by voting from 1-star (very bad) to 5-stars (very good) for each dialogue.

The results of human evaluation on the baseline dialogue models, shown in Table 2.5, indicate that 91% of the system turns retrieved by TF-IDF were considered appro-

priate and coherent by the test users. As a result, more than 70% of the dialogues had acceptable quality. These results suggest the usefulness and suitability of the elicited dialogues for longitudinal and multi-session conversations. Meanwhile, even though BM25 managed to outperform the TF-IDF in automatic evaluation, TF-IDF obtained higher ratings both at the turn level and at the dialogue level.

2.5 Case Study: Personal Healthcare Agent

In collaboration with a team of psychotherapists, we developed a Personal Healthcare Agent (PHA) for the mental health domain using the collected dataset of LDs[3]. The PHA was embodied in a mobile application available to users in both Android and iOS platforms, and was capable of engaging the users in two types of dialogues with the goal of improving their mental health.

In the first interaction type, the users could initiate a dialogue with the PHA At any time of the day about a real-life event that has activated their emotional state. In this case, the PHA would engage the user in a dialogue by asking a controlled set of questions designed by the therapists to obtain more details about the event using the ABC technique. The answers of the user were then presented to the therapists to provide support to the patient accordingly.

Regarding the second interaction type, the PHA would hold a system-initiated dialogue asking about how the user feels about the event and the emotions she shared the day before. The PHA would create a personal dialogue stimulus for the user, and engage the user in a dialogue to follow-up with them and check whether the issue is fully resolved or more therapeutic support is required. During this interaction, the PHA would provide helpful suggestions during the conversation with the user and support him/her to reach a healthier emotional state. These conversations consisted of natural language responses from the users and retrieved responses from the system using BM25. During each follow-up interaction, the model would select the top 3 response candidates. Afterward, it would rank the candidates based on their coherence according to the recurring entities that appeared in the dialogue history. The most appropriate response would be selected and output to the user. the developed PHA was deployed in two clinical pilot studies.

2.5.1 Pilot Study 1: Participatory Design

In our first study, the participants and the psychotherapists were engaged in the early phases of the design and development of the application. 21 participants aged 33-61 with mild-to-moderate levels of stress, anxiety and depression were assigned to two groups, A and B. While both groups received stress management training sessions along with cognitive behavioral treatment, Group A interacted with the PHA as well. Psychopathological outcomes were assessed at baseline (T1), following eight weeks of treatment (T2), and after three months post-treatment (T3).

[3]The protocol and the experimental plan were approved by the Ethical Committee of the University of Trento, Italy. Our experimental protocol has been registered on ClinicalTrials.gov (NCT04809090).

The results of this study supported the hypothesis that stress management treatment can benefit from the deployment of personal dialogue systems, in particular for improving adherence to therapists' recommendations about applying coping strategies in everyday life. Those improvements have been shown to persist over time. The patient group who interacted with the PHA reported significant improvements between T1 and T3. Moreover, the psychotherapists engaged in this study were in favor of integrating a PHA into their practice as they could observe increased engagement of patients in pursuing therapy goals. Further details about this study are published in our article [10].

2.5.2 Pilot Study 2: Randomized Controlled Trial

In our second study, we evaluated the contribution of our PHA in promoting mental health and well-being. This study was based on a protocolized intervention for stress and anxiety management where patients with stress symptoms and mild-to-moderate levels of anxiety received eight weeks of CBT treatment delivered remotely. The participants were active workers aged over 55. Four experimental groups were selected; G1 received traditional therapy, G2 also conversed with the agent, G3 received support only by the agent, G4 did not receive any treatment and was assigned to a waiting list. The symptoms related to stress were assessed prior to the treatment (T1), at the end (T2), and three months after (T3) by standardized psychological questionnaires.

Analysis conducted within groups showed greater improvements in the levels of stress and scales related to overall well-being in G2. Besides, G2 reported higher levels of perceived usefulness and satisfaction. Moreover, we observed a greater level of satisfaction and subjective perception of usefulness in participants who could be supported by the human therapist as well as the PHA. Further details about this study are published in our article [11].

2.6 Conclusions

We addressed the need for suitable dialogue corpora for Longitudinal Dialogues (LDs) by presenting an elicitation methodology for LDs. Using the proposed methodology, we collected a dataset of LDs consisting of 800 2-session dialogues in the mental health domain.

Through an analysis of the collected corpus 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.

We investigated the appropriateness of the collected corpus for developing multi-session dialogue systems. 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 91% of the system turns as appropriate and coherent, resulting in 72% of dialogues with acceptable quality.

CHAPTER 3

Personal Knowledge Extraction

The knowledge required to carry out Longitudinal Dialogues (LD) is user-specific and can vary for each dialogue session with the individual user. There is no general-purpose knowledge base that can suit all users and scenarios.

User responses in LDs have a unique and complex structure. They encompass personal events and situations the user has experienced and shares with the machine. We present an unsupervised model to automatically parse the user response and extract the user's personal events and participants as personal knowledge. This information is then presented as a graph of the user's personal space. This personal graph is then updated at each interaction with the patient. The obtained graph is further used as a source of knowledge for grounded response generation in LDs.

3.1 Background

The definition of the event concept has been the topic of study in different disciplines, originating in philosophy [47]. Early linguistic attempts to understand the semantics and structures of events in unstructured text date back to the use of hand-coded scripts (frames) [32]. In this approach, predefined slot frames were designed to be filled by the values extracted from the text. This approach was later adopted by Ebner et al. [14] where the authors studied the events and their participants by the verb-specific roles the participants can have (the arguments of the event "attack" are of types "attacker" and "target"). In this work, the authors formalized the event understanding as an argument-linking task. Kim and Klinger [30] consider the activation of emotions as an event and study such events through different properties such as cause, experiencer, target, etc. In

this definition, not only verb phrases but also noun phrases and prepositional phrases that activate an emotion in a participant can represent events.

In order to address the expensive nature of designing domain-specific frames, Chambers and Jurafsky [5] proposed an unsupervised approach to extract the event chains in a narrative according to the common protagonist they surround. Based on the assumption that reoccurring participants among different events are the protagonists of the narrative, the authors defined the event in a sentence by its predicate (verb) and verb dependencies. This work was complemented further by considering the role of the protagonists in each event and the neighboring events in order to obtain a schema [6].

There have been several studies on the application of narrative understanding through event extraction and annotation. In this regard, Mostafazadeh et al. [46] applied event chain extraction model [5] for the task of closure selection for commonsense stories, known as StoryClozeTest. Rashkin et al. [63] conducted a task on inferring the next possible intents and reactions of the participants in a narrative based on the observed events through commonsense. Zhou et al. [100] studied the application of temporal reasoning such as the order/frequency of events in the narrative for the question-answering setting.

3.2 Personal Space Graph

We present an unsupervised approach to automatically extract the life events and their participants from the user responses, and represent them as the Personal Space Graph (PSG) of the user. We follow the definition of an event that was used by Chambers and Jurafsky [5] based on the verb and its dependencies. That is, a verb is the core element of an event and supports the relation among its dependencies such as subject, object/oblique nominals which are considered as the participants of the event.

Figure 3.1 shows the workflow of our model. Through the interaction with the user, each response is parsed and presented in terms of its predicates (the events, the edges of the graph) and their noun dependencies (the participants, the nodes of the graph). Each edge has an index based on its order of appearance in the narrative which makes it possible to reconstruct the order of occurrences among the events (for instance, the event "*litigo spesso* (argue often)" is mentioned after "*parla male* (talks bad)"). Besides, the events and participants mentioned in a recent dialogue session are considered to be more relevant for ongoing interaction. Based on this assumption, older nodes and edges in the graph will become less relevant upon a new dialogue (presented by dashed lines in Figure 3.1).

Our architecture consists of five main components:

1. **Functional Unit Segmentor** Upon receiving a response, it is first segmented into its functional units. A functional unit is a contiguous span within a message which has a coherent communicative intention [54]. The segmentation into functional units was performed by a seq2seq model presented in Chapter 2, Table 2.3.

2. **Dependency Parser** Each functional unit is then passed to the dependency parser

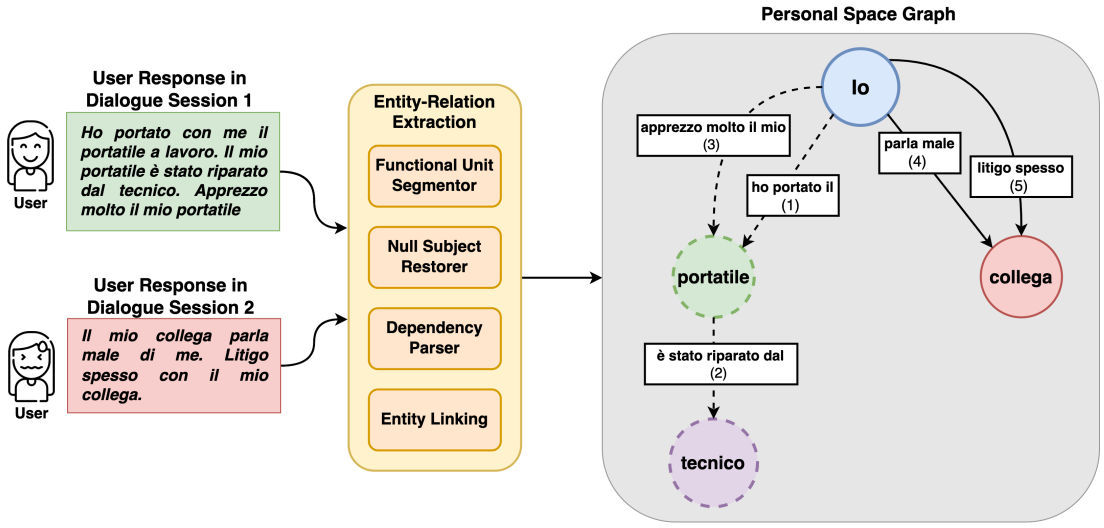

Figure 3.1: The events and participants in the user responses are extracted and presented as the Personal Space Graph (PSG) of the user. Each edge and the adjacent nodes stand for an event and its corresponding participants, respectively. The edges (events) have an index based on their appearance in the response. Events and participants extracted from prior dialogue sessions are considered less relevant for ongoing interaction and have a lower importance score, presented by dashed lines.

to obtain the corresponding dependency tree, for which spaCy natural language processing library[1] was used. Using the obtained tree and part-of-speech tags, tokens tagged as nouns and proper nouns are extracted as nodes in the graph (nominal modifier nouns are excluded in this process since they are describing/specifying characteristics of another noun). In cases where pronouns are subjects or objects of a verb, they are extracted as nodes as well.

3. **Entity Linking** In order to make sure different surface representations of the same noun are mapped to the correct node in the graph, an Entity Linking module is defined. This module queries BabelNet[2] and ConceptNet[3] semantic networks for the root form of the extracted nouns and matches them consequently to obtain a set of entities and participants in the narrative.

4. **Null Subject Restorer** All the verbs contained in the functional unit are extracted and controlled for possible null subject cases. Null subjects are non-overtly expressed subject pronouns commonly used in pro-drop languages such as Italian and Spanish [69]. In this case, the subject of the verb is restored as a pronoun based on its conjugation using an out-of-the-shelf library MLCONJUG3[4] to make sure each participant is detected and extracted correctly.

5. **Entity-Relation Extraction** Lastly, the model navigates through the dependency

[1] spaCy `spacy.io`
[2] BabelNet `babelnet.org`
[3] ConceptNet `conceptnet.io`
[4] MLCONJUG3 `pypi.org/project/mlconjug3`

tree to find the verbs that connect the extracted entities as subjects and objects/oblique nominals. In cases of entity conjunctions, the same verb spans over all the entities in the same conjunction. For better visualization, the neighbors of the verb in the dependency tree are explored to obtain an entire predicate composed of adverbs, ad-positions, and auxiliaries as the edge of the graph.

The obtained PSG is specific to each user and presents the mentioned event and participants in the narratives. In each graph, the user is presented as the node "*Io* (I)" and all the other participants are connected to it by the corresponding predicate.

3.3 Evaluation

We evaluate our proposed approach in two different settings in the Italian language. Besides, we compare the performance of its English adaptation with other models in the StoryClozeTest [46] setting.

3.3.1 Italian Corpus

For the evaluation of the proposed model, we collected a dataset of human-machine dialogues using the approach introduced priorly in Chapter 2 for the collection of first-session dialogues. The users were asked to interact with the dialogue agent and answer a set of questions about real-life situations and events that have activated their emotional state for the period of three months. As the result, we collected 241 dialogues from 18 users with an average length of 128.2 tokens per dialogue (excluding the machine turns) and an average number of 11.9 dialogues per user.

In the first setting, we evaluated the model for the task of last-turn selection in a personal dialogue. Using the collected dialogues, the model was tasked to select the last user response in each dialogue based on the participants and events (verbs) it consists of and the PSG it extracted from the dialogue history. We assessed the performance of the model using two pools of 2 and 5 candidates, each consisting of 1 correct response and *n-1* distractors. The distractors were sampled randomly from the pool of the last user turns in other dialogues in the same dataset.

In the second setting, we evaluated whether the obtained graph can correctly represent a personal space of events and participants that varies for each user. The model was first presented with a set of consecutive dialogues from a specific user as history. It was then tasked to select the next possible dialogue for the user based on the PSG extracted from the previous dialogues. We evaluated the mode using a pool of 2 candidates, consisting of the correct next dialogue and a distractor (a dialogue with a different user.)

The results of these evaluations are presented in Table 3.1. In the first scenario, while TF-IDF manages to be a strong baseline, our proposed system outperforms the Random baseline and has a higher success rate than the selection solely based on the recurrence of the nouns. By raising the task difficulty and increasing the pool size to 5, our model maintains the same performance trend. Regarding the second evaluation, the results indicate that the recurrence of the nouns is an important factor for the model to select the next possible dialogue. Nevertheless, our model manages to outperform

Last User Turn Selection					Next Dialogue Selection (Pool of 2)				
Recall	**Rand.**	**TF-IDF**	**Nouns**	**PSG**	**History**	**Rand.**	**TF-IDF**	**Nouns**	**PSG**
R@1 in 2	50%	71.1%	41.3%	59.0%	**2 Personal Dialogues**	50%	74.4%	68.8%	71.4%
R@1 in 5	20%	51.6%	34.8%	42.7%	**5 Personal Dialogues**	50%	75.3%	68.8%	72.0%

Table 3.1: The results of evaluating our model in the Italian language in two different settings.

	EC	**Nouns**	**PSG**
R@1 in 2	49.4	45.1	45.6

Table 3.2: The result of evaluating the English adaptation of our model in StoryClozeTest setting, compared with other unsupervised approaches [46]. EC stands for Event Chain model baseline [5].

this baseline by considering the predicates as an additional factor, and gets closer to TF-IDF scores.

3.3.2 English Corpus

The English adaption of the model was evaluated in the StoryClozeTest setting. In this setting, the model is tasked to select the most probable ending for a four-sentence story from a pool of 2, consisting of the right ending and the wrong one [46]. We compare the performance of our model with the Event Chain (EC) model proposed by Chamber and Jurafsky [5] that follows the same linguistic definition for an event. The result of this evaluation for the test set of 3744 stories is presented in Table 3.2, indicating that our model performance is in line with other unsupervised approaches.

3.4 New Event Detection

Throughout our analysis of the PSG model performance, we observed that one of the main sub-challenges in extracting the PSG of the user is identifying the novelty of the extracted events. To obtain a concise and salient understanding of a narrative through the events, it is necessary to select the events that relate to a new happening/participant in the narrative and have novel contributions. The process of recognizing an event new implicitly involves the event coreference resolution task. This task consists of detecting all the events that refer to the same event [94]. Thus, an event that is referring to a previous event is not considered new. Nevertheless, if the event appears in the narrative for the first time it might be part of commonsense knowledge and thus not new.

We assess whether an event is new in a narrative according to their Information Status (IS) [42,58]. This study is inspired and motivated by the need to a) extract salient information in the narrative and position them with respect to the rest of the discourse events and relations, and b) acquire a new event from a sequence of sentential units of narratives. This task can facilitate higher levels of computation and interaction such as reasoning, summarization, and human-machine dialogue. We annotated a publicly available corpus of narratives with the new events at the sentence level using human

annotators. We then developed several neural and non-neural baselines for the task of new event extraction in both candidate-selection and sequence-tagging settings.

3.4.1 Definition of New Event

Prince [58] defined the notion of old or new Information Status (IS) with respect to two aspects of the hearer's beliefs and the discourse model. New information according to the hearer's belief is the one that is assumed not to be already known for the hearer, while discourse-new information is the one that has not been mentioned or has not occurred priorly in the discourse-stretch [58]. Nissim et al. [52] adopts the IS concept and defines three categories of old, new, and mediated for the status of entities in a dialogue. The notion of old follows the definition provided by Prince [58] closely. However, the authors define mediated as entities that have not directly been introduced in the context but are inferrable or generally known to the hearer; while the new category spans over entities that are not introduced priorly in the dialogue context, nor can they be inferred from the previously mentioned entities.

So uh during my childhood **I had two dogs; one was named Flash**, **one was named Fluff**.

I got them when I was three and around the age of eight, **we were moving to the US** from Guyana.

When we were living in the US, **we rented a house** for a short time and **my father bought a big sofa.**

Figure 3.2: Sentences in a narrative and the corresponding events. There are eight events in the sentences (highlighted), while six of them are presenting new information (bold) and the remaining two are referring to the already-mentioned events in the context (not bold).

We extend the definition of the new category in entities [52] to events. We define an event as new if its information (the event and/or participants) is not presented priorly in the discourse stretch, and it can not be inferred through commonsense. For instance, *Bob **saw** Alice* is a new event if it is the first time that Alice is introduced in the narrative or the first time Bob saw her. However, once this event is selected as new, *Bob **looked at** Alice* will not be a new event anymore. Furthermore, if *Bob **married** Alice* is considered as a new event, *Alice **is** Bob's wife* can be inferred through commonsense and thus is not a new event.

An example of new and old events is presented in Figure 3.2. While there are eight events in the narrative sentences, two of them do not represent any novel information and thus are not new.

3.4.2 Annotation of New Event

We conducted an annotation task for identifying the new events in narratives at the sentence level. The corpus used in this study is the SEND [55]. This dataset is a collection of personal emotional narratives, collected by asking each subjects to recount 3 most positive and 3 most negative experiences of her/his life. This property makes this dataset an appropriate corpus to study personal narratives about events and the emotional activation of the narrators. The dataset consists of 193 narratives from 49

	Value
#Narratives (Train:Valid:Test)	193 (114:40:39)
#Subject (# female)	49 (30)
Avg. Narrative Len.	28.10 utterances
Avg. Utterance Len.	15.44 tokens
#Vocabulary	4,416 unique tokens

Table 3.3: The statistics of SEND dataset [55]. The dataset is provided with official train, valid, and test sets. Each narrative consists of approximately 430 tokens on average.

subjects. The statistics of the SEND dataset are presented in Table 3.3 (the train, valid, and test sets are the official splits).

To reduce the annotators' workload, we used the English adaptation of our PSG framework to automatically parse and extract all event candidates for each sentence in the narrative as the triplets of (subject, predicate, object). In the cases where more than 5 candidates were extracted for a sentence, we created 5 clusters using Levenshtein distance [92] (hierarchical clustering) and the candidate with the most number of tokens in each cluster was selected to be presented to the annotator. We randomly sampled 21 narratives from the SEND dataset and reserved them as backup data (13 narratives from the train set, 4 from the valid set, and 4 from the test set). Using the extraction pipeline, we extracted all subject-predicate-object triplets as event candidates in the remaining 172 narratives at the sentence level.

We recruited five annotators. During the task, the annotators were presented with a narrative one sentence at a time and the corresponding list of candidates. They were asked to control if any of the candidate triplets in the list is valid (i.e. it reflects the information in the sentence correctly); and whether it provides new information with respect to the previous narrative context, that can not be inferred through commonsense. In the case of valid and new information, the annotators were asked to select that candidate as a new event. Furthermore, if there were no candidates extracted for a sentence or the new information in a sentence was not presented as a valid candidate, the annotator was asked to add the new information by simply copying the segment that conveys it from the sentence and adding it as continuous span text.

After an introductory meeting with the annotators, they were asked to carry out the first qualification task which consisted of annotating one narrative, sampled from the valid set. The result of the first qualification batch was checked manually and a few refinements were made with the annotators. The annotators were then asked to perform a second qualification task using another narrative randomly sampled from the valid set. The Inter-Annotator Agreement (IAA) level during the two qualification tasks, which is presented in Table 3.4, indicates the improvement in the annotators' performance from one qualification batch to the other. The IAA for the event candidates is calculated using Krippendoff's α [31], while the IAA for the continuous span text is calculated by the extension of Cohen's κ for segmentation agreement [17], averaged among all annotators.

The remaining 170 narratives were divided into 11 batches. In each batch, one narrative was annotated by all annotators for the purpose of continuous quality control

Annotation Format	Qualifications First	Qualifications Second	Overall IAA in the Annotation Task
Selected Candidates	0.22	0.55	0.54
Added Continuous Spans	0.32	0.60	0.66

Table 3.4: Inter-Annotator Agreement (IAA) during the qualification tasks and over the whole annotation task. The results indicate an improvement in the performance of annotators from one qualification batch to the other. The IAA is computed for candidate selection and continuous span selection annotation using Krippendoff's α and the extension of Cohen's κ for segmentation agreement, respectively.

of the results, while the rest was equally divided among the annotators. To prevent unreliable and biased agreements, all 11 overlapping narratives were from different narrators.

3.4.3 Evaluation of Annotated Corpus

Throughout the task, the IAA level on the overlapping narratives was computed to ensure a consistent annotation quality. We observed negligible fluctuations in the IAA level during the task (<0.9 for Krippendoff's α), except for one batch; for which the low-quality contributions were detected and refinements were made with one annotator. The overall IAA level of the annotated dataset is presented in Table 3.4. The results are close to the level obtained in the second qualification batch.

Sentence 1: So uh during my childhood I had two dogs; one was named Flash, one was named Fluff.

Candidates:
a. [i] - [had] -> [my childhood]
b. **[i] - [had] -> [two dogs]** ✔
c. **[one] - [was named] -> [fluff]** ✔
d. [i] - [so had] -> [two dogs]
e. **[one] - [was named] -> [flash]** ✔

Sentence 2: I got them when I was three and around the age of eight we were moving to the US from Guyana.

Candidates:
a. [i] - [got] -> [them]
b. **[we] - [were moving tol -> [the us]** ✔
c. [we] - [were moving to] -> [guyana]

Sentence 3: When we were living in the US, we rented a house for a short time and my father bought a big sofa.

Candidates:
a. [we] - [were living in] -> [the us]
b. **[we] - [rented] -> [a house]** ✔

Added Spans: ***my father bought a big sofa***

Figure 3.3: Sentences in a narrative and the corresponding events; while the baseline model has extracted various event candidates, only a few of them are valid and new events (bold). Furthermore, the baseline model has missed an event in the third sentence which is added as a span from the sentence.

The results of the annotated dataset, presented in Table 3.5, indicate that the majority of the annotated events were added as continuous span text and were not extracted by the PSG model. An example of the annotation results is presented in Figure 3.3. While the event candidates appear in the narrative with an approximately uniform distribution, almost all of the continuous span events are located in the first half of the narrative. This result is in line with the definition of new events since the events mentioned before in the context are "old" events. Nevertheless, in both cases of candidate events and continuous span events, we observe that the sec-

Selected New Events as Candidates	
#Candidates selected	1536
Avg. candidates selected:	
per Sentence	0.57
per Narrative	9.0
per Narrator	31.4
%Candidates selected in:	
1st half of the Sentence	43%
2nd half of the Sentence	57%
1st half of the Narrative	55%
2nd half of the Narrative	45%
Added New Events as Continuous Spans	
#Spans added	2254
Avg. spans added:	
per Sentence	0.8
per Narrative	13.3
per Narrator	46.0
%Spans added in:	
1st half of the Sentence	38.1%
2nd half of the Sentence	61.9%
1st half of the Narrative	96.9%
2nd half of the Narrative	3.1%

Table 3.5: The statistics of the annotated dataset. While only 1536 extracted candidates (out of 6938, thus 22%) were selected as new events, 2254 new events were added by the annotators as continuous span text. Moreover, almost all of the continuous span events appear in the first half of the narrative, while event candidates have a quite normal distribution.

ond half of the sentences contains more information than the other half, indicating that the narrators tend to mention the new events at the end of the sentence.

3.4.4 Baselines for New Event Detection

We developed neural and non-neural baselines to validate the outcome of the annotation task, and, as baselines for the novel task of new event detection in a narrative. Considering the two annotation formats of selecting candidates and adding continuous spans, we formalize the task using two settings of candidate selection and sequence tagging.

A) Candidate Selection Baselines

The first group of models is tasked to select the new events from the candidates extracted by our PSG model. The rule-based models are:

- **Random Selector**: for each sentence and its event candidates, it randomly picks one candidate as the new event in the sentence.

3.4. New Event Detection

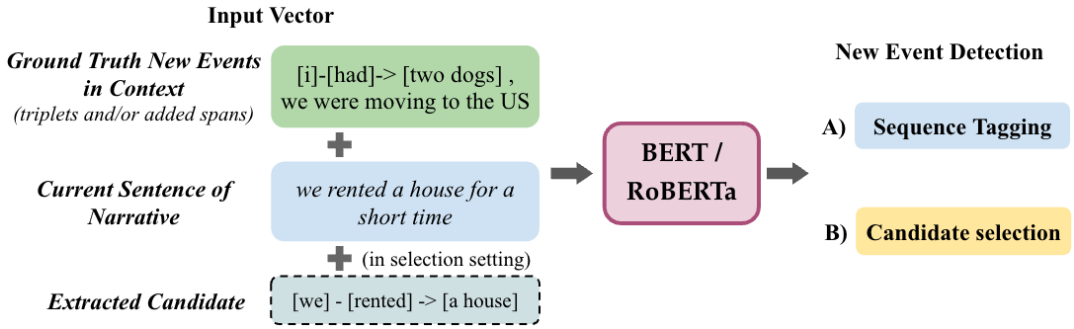

Figure 3.4: The neural baselines for the task of new event detection. The input vector consists of the new events in the context (ground truth) and the current sentence. In the candidate selection setting, the input vector includes the extracted candidate as an additional segment as well. The model encodes the input vector and outputs either a) a sequence of tags, corresponding to the tokens in the sentence; or b) a binary decision to categorize the candidate as new or not.

- **Binary Selector**: for each of the event candidates of a sentence, it randomly decides whether it is a new event or not. Thus, each candidate has a 50% chance of being selected as a new event.
- **First Candidate Selector**: that selects the first event candidate that is extracted for a sentence as the new event.
- **Last Candidate Selector**: which selects the last event candidate that is extracted for a sentence as the new event for the sentence.
- **New Subject Selector**: which selects the first candidate that contains a new (unseen) subject in the list of candidates as the new event. In other words, the number of selected candidates is equal to the number of non-repetitive subjects in the candidate list of the narrative.
- **New Entity Selector**: which selects all the event candidates that include new subjects or new objects at the narrative level. Thus, it selects all candidates unless they differ in the verb only. In that case, it selects one of them as the new event.

Neural Network Models In addition to the rule-based models, we developed neural models based on Pre-trained Language Models (PLMs) as baselines for the task of new event candidate selection presented in Figure 3.4. For this purpose, we model the input vector with three elements as event candidate, current sentence, and context new events. The context new events denote the new events (ground truth) in the narrative context up to the current sentence. In cases where the size of the input vector exceeds the model limits (for instance 512 tokens per BERT-based models), the model trims the former part of the context new events. The model encodes this vector and outputs the classification decision of whether the event candidate (triplet) is a new event or not. The PLMs we fine-tuned for this purpose are BERT [13], and RoBERTa [38].

The results of the candidate selection baselines are presented in Table 3.6. We observe that *Last Candidate Selector* has achieved the highest precision level among

	Prec.	Rec.	F1
Random	24.0	29.2	26.3
Binary	22.8	49.4	31.2
First Candidate	27.7	33.7	30.4
Last Candidate	30.1	36.7	33.1
New Subject	24.6	28.6	26.5
New Entity	25.1	88.9	39.1
BERT	35.6	51.1	41.6
RoBERTa	40.4	83.1	**54.3**

Table 3.6: The results of the new event candidate selection baselines. The performance of the neural models is averaged over 10 runs.

	Prec. (%)	Rec. (%)	F1 (%)
Random	18.8	49.7	27.3
Early	17.4	29.5	21.9
Late	20.2	34.0	25.4
BERT	33.2	82.2	47.3
RoBERTa	34.3	81.3	**48.3**

Table 3.7: The results of the new event sequence tagging baselines. The models are trained and tested on continuous span events annotated by the human judges only. The performance of the neural models is averaged over 10 runs.

	Prec. (%)	Rec. (%)	F1 (%)
Random	31.1	49.6	38.2
Early	30.8	31.6	31.2
Late	29.9	30.4	30.2
BERT	54.9	84.3	66.5
RoBERTa	55.5	84.8	**67.1**

Table 3.8: The results of the new event sequence tagging baselines. Compared to Table 3.7, in this setting, the models are trained and tested on both selected candidates and continuous span events annotated by the human judges. The performance of the neural models is averaged over 10 runs.

rule-based models. This is in line with the annotation result analysis, indicating the percentage of selected new event candidates to be slightly higher at the end of sentences. On the other hand, *New Entity Selector* achieves the highest level of recall while having a very low level of precision, as it selects all candidates unless the variation is only in the verb predicate. Moreover, the F1 scores of all the rule-based models are less than 40.0%. This indicates that features such as the novelty in elements or occurrence position are not enough to achieve high performance on the task of new event selection. While both neural models outperform the rule-based ones, RoBERTa outperforms all the baselines in this task by having the highest level of precision while maintaining a high recall.

B) Sequence Tagging Baselines

The second group of the models is developed for the task of new event detection in a sequence tagging setting. That is, the models tag the sequence of tokens (chunks) which are representing a new event in the sentence. The analysis performed on the continuous span events selected by the human judges indicated that several events can

share the same tag spans such as subject or object. Therefore, we formalize this task as a binary tagging task rather than IOB tagging task and leave the development of the models for IOB tagging of multiple spans with overlap as future work. Similar to the previous task, we developed rule-based and neural baselines for new event sequence tagging. The developed rule-based baselines are:

- **Random Tagger**: which randomly tags tokens in a sentence as the new event tokens.
- **Early Tagger**: which tags the tokens in the first 30% of a sentence as the new event tokens.
- **Late Tagger**: which tags the tokens in the last 30% of a sentence as the new event tokens.

Neural Network Models Using BERT [13], and RoBERTa [38] PLMs, we developed two neural baselines for this task. The models take as input the current sentence and the context new events which are the sequences of new events in the narrative context up to the current sentence. Similarly to the previous neural baselines, if the input vector exceeds the size limits of the models the former part of the context new events is trimmed. The model encodes this vector and outputs a tag sequence consisting of $\mathbf{E}_{(\text{vent})}$ or **O**, corresponding to the tokens in the sentence, indicating whether or not they describe a new event.

We initially trained the sequence tagging baselines using the annotated continuous span events. The results of this experiment are presented in Table 3.7. We observed that precision scores and consequently F1 scores are not significantly different among rule-based models. This indicates that the position of the tokens in the sentence is not the most contributing factor to the prediction accuracy. Similar to the previous task, the neural models have the highest performance among the baselines. However, their precision is considerably lower than the recall.

In the next step, we evaluated the same baseline models using both the selected event candidates and the continuous span annotations as the train and test sets. The results of this experiment, presented in Table 3.8, show a boost in the performance of all models using the mentioned train and test sets. Nevertheless, the same performance trends among models can be observed in this experiment as well.

3.5 Conclusion

In this work, we present an approach to automatically extract life-events and participants from user responses in Longitudinal Dialogues (LD) and represent them as a personal graph. This graph can be a source of knowledge for dialogue systems designed for LD.

To identify the events that present novel information in the narrative, we study the unfolding of the events according to their Information Status. We introduce the new task of identifying new events as they unfold in the narrative. We annotated a complete dataset of personal narratives with new events at the sentence level using human

annotators. We then developed several neural and non-neural baselines for the task of new event detection in both settings of candidate selection and sequence tagging. We believe this task can be a novel and challenging task in natural language processing and can support other tasks in human-machine dialogue and natural language generation.

CHAPTER 4

Human Evaluation Protocol

Human Evaluation (HE) of automatically generated responses is necessary for the extrinsic evaluation of the human-machine dialogue systems. Early attempts to evaluate automatic Natural Language Generation (NLG) models using human judges date back to the 90s, before the appearance of end-to-end models [7, 26, 33]. However, due to the expensive requirements such as training skilled annotators and the time-consuming nature of this evaluation, automatic metrics became the common evaluation criteria in several NLG tasks. Metrics such as BLEU [56], METEOR [2] and ROUGE [36] have been used to evaluate the model performance in machine translation and automatic summarization tasks respectively as inexpensive and rapid evaluations. After observing the reliability of these metrics for the task they are designed for (if applied correctly), they have been used to evaluate the models in other tasks such as response generation. However, several studies have shown that such metrics can not be reliable proxies for evaluating generated responses [37, 70]; these criteria co-relate poorly with human judgment and are inadequate since the generation is subject to linguistic features such as grammaticality, fluency, and coherence, as well as interaction features such as appropriateness, engagement, and user acceptance.

With the development of crowd-sourcing annotation platforms, conducting an HE task is less expensive and more feasible than early methodologies. Nonetheless, due to the lack of agreed-upon HE protocols, numerous HE tasks have been introduced in the community suffering from incomparable and different characteristics, nontransparent procedures, and non-replicable and incomparable results.

We propose to standardize the experimental methodology for HE of response generation models. We present a detailed protocol for this task, in order to increase the comparability, replicability, and interpretability of such evaluations among works and do-

mains. We present all the required steps and materials to conduct a HE in a transparent and extendable way. The proposed protocol is domain-agnostic, language-independent, and open to being extended to different versions and standards.

4.1 Background

Earlier attempts to evaluate dialogue systems by human judges considered user satisfaction as the evaluation criterion [81]. Despite the introduction of automatic metrics for the evaluation of dialogue models and a research direction aiming to better the metrics used [23, 44, 96], HE is still the gold standard for assessing the qualities of a generated response and a generative model [75].

While the importance of the proper evaluation of a dialogue model using human judges is well-established in the community, how to perform such evaluation is still an unsolved question [75]. As an outcome, countless HE tasks have been presented and conducted in this domain, resulting in non-comparable and non-replicable results. Dialogue systems have been evaluated with different granularity (turn-level vs. dialogue-level), different evaluation policies (single-model vs pairwise-model, candidate-ranking vs. winner-selection), and in different modalities (interactive vs. static) [75]. The ambiguities in HE tasks conducted so far have also been studied by Belz et al. [3], where the authors focused on disentangling the characteristics of already conducted HE tasks to increase the interpretability and comparability of the evaluations and results. Further inconsistency in the evaluations includes the ambiguity in the criterion name, i.e. two criteria with the same name assess two different qualities in different works, whereas the same quality has been named with various terms among works [22]. In addition to the aforementioned works, this naming inconsistency can also be found in the grounded generation literature [21, 25, 83, 97] where a criterion with the same name refers to two different qualities and presents different definitions among works.

An important factor for reproducing any crowd-sourcing experiment is reporting the details related to that experiment and its settings. This issue has been studied by Ramirez et al. [61], where the authors identify the properties that researchers have to provide to facilitate the reproducibility of any crowd-sourcing experiments. The same problem has been studied specifically for HE experiments by Howcroft et al. [22] where the authors identify the lack of reporting crucial details and other issues such as high levels of variation among the evaluation procedures. Howcroft et al. [22] further stress the need for a standard and coherent experimental design and terminology for the task of HE in the community.

4.2 The HE Annotation Protocol

We propose to standardize the HE experiments through a referable and replicable protocol[1] to address the problems of non-comparability and inconsistency in the literature. Figure 4.1 presents the diagram of our proposed protocol. Considering the complexity

[1] https://github.com/sislab-unitn/Human-Evaluation-Protocol

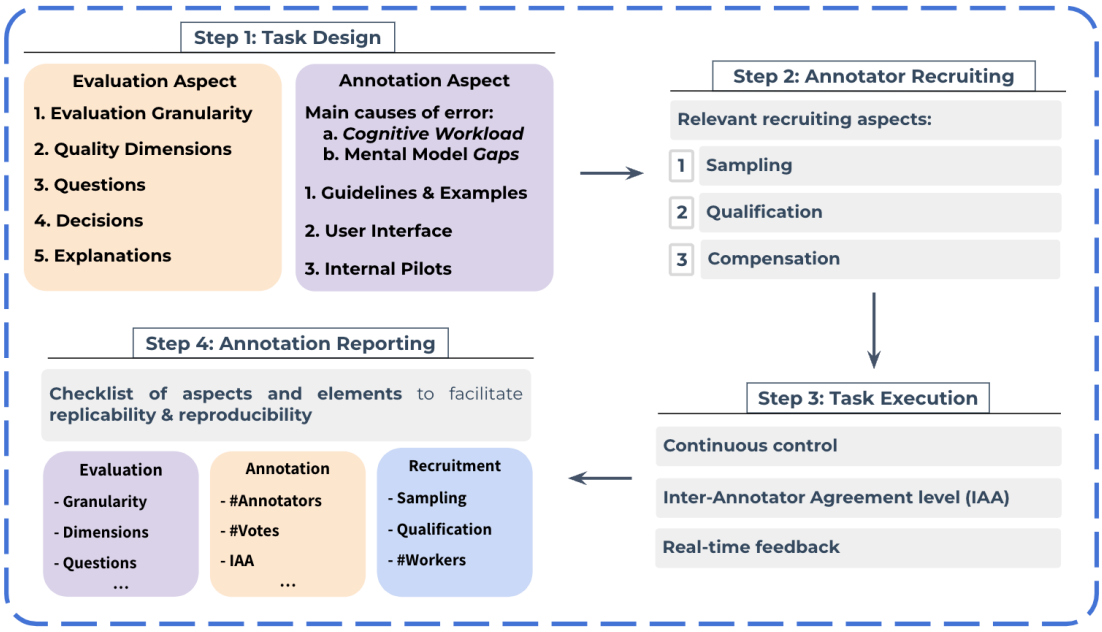

Figure 4.1: Our proposed Human Evaluation protocol consists of four main executive steps. In the first step, we unfold the Task Design according to the evaluation and the annotation characteristics. We then present the necessary factors to recruit a suitable group of annotators (crowd-workers) in the second step. Step 3 includes the required considerations during the execution of the task to ensure a high-quality outcome; while the last step provides a guideline for crucial information to report in order to present a transparent and replicable evaluation.

of designing and executing such evaluations, we unfold the task into four main steps in order to study and analyze the crucial aspects at each step. We aim to maximize the reliability and replicability of the evaluation while minimizing the task difficulty and complexity.

Our proposed protocol consists of four executive steps, i.e. 1) Task Design; 2) Annotator Recruiting; 3) Task Execution; and 4) Annotation Reporting. In the first step, we study the required characteristics to design a reliable and replicable evaluation task with respect to both the evaluation aspect and the annotation aspect. The second step is dedicated to identifying the crucial factors for recruiting an appropriate group of annotators to perform the task; while the third step presents the required actions to monitor and ensure annotation quality. Lastly, we present a complete checklist of the necessary information to report alongside the evaluation results to achieve replicability and transparency of the HE tasks.

4.2.1 Task Design

The first step is to design the evaluation task, which can be characterized by the two aspects of evaluation and annotation. Defining these characteristics clearly and transparently is paramount to achieving replicability and comparability among works and models.

A) Evaluation Characteristics

As the initial step, the definition of the evaluation characteristics of the task includes the evaluation granularity, quality dimensions to evaluate and their definitions, the questions to be asked to the annotators, and the annotations format.

Granularity The evaluations conducted in the literature can be categorized into two levels of granularity as dialogue-level, where the model is evaluated at the end of a complete dialogue, and turn-level, where the model is evaluated based on its output for a specific turn in the dialogue. Recent works indicate that the turn-level evaluation is more fine-grained since it captures errors such as contradictions and response repetitions [75]. Turn-level evaluation can be further categorized as absolute (single-model, or rating) or comparative (winner-selecting, or ranking). In this protocol, we evaluate the models at the turn-level; and in order to avoid biasing the annotators with the quality of other candidates which may result in an unintentional pick-the-best response, we evaluate the candidates using the absolute setting (i.e. presenting one candidate per time for each dialogue history). In this way, the performance quality of each model is evaluated independently and we can obtain a model-specific list of limitations and error signals. Furthermore, the ground truth turn is also provided as a response candidate to the annotators, representing a point of reference.

Quality Dimensions We include four criteria in this version of the protocol, based on the most common errors and qualities for an end-to-end response generation model. Nevertheless, the proposed protocol can be extended to other criteria and quality dimensions. The proposed criteria and their definitions are as follows;

- **Appropriate** whether the proposed response candidate makes sense with respect to the dialogue history; and to investigate if it is a proper continuation of the given dialogue (thus coherent).

- **Contextual** whether the proposed response candidate contains references to the dialogue context (thus not generic); and to investigate whether the response refers to non-existing or contradicting information (such as model hallucination).

- **Listening** whether the speaker of the proposed response is following the dialogue with attention (note that generic responses are also indicating that the speaker is not following the dialogue).

- **Correct** whether the response candidate is correct considering the grammar, syntax, and structure of the response.

Questions One of the important details, which is usually missing in the evaluation reports in the literature, is the formulation of the questions the annotators are prompted for the quality of the responses. The questions must be designed in a clear and neutral form in order to avoid any possible bias while addressing the important factors evaluated by each criterion. We present the questions designed to evaluate the responses in each dimension in Table 4.1 (The protocol can be expanded to other dimensions used by adding the corresponding criteria and questions).

Dimension	Question	Answer Option	Option Definition
Appropriateness	*Is the proposed response candidate appropriate?*	Appropriate	The response makes sense and it can be the natural continuation of the shown dialogue context.
		Not Appropriate	The response does not make sense in the current dialogue context.
		I don't know	The candidate contains some elements which make sense with respect to the dialogue context, but some that do not.
Contextualization	*Does the proposed response contain references to the context of the dialogue?*	Contextualized	The candidate contains implicit or explicit references to the dialogue context.
		Not Contextualized	The candidate doesn't contain any reference to the dialogue context, or contains references that are incoherent with the dialogue context.
		I don't know	The response contains some references to the dialogue context, but contains other references that are not clear or relevant.
Listening	*In the proposed response candidate, how much do you think person A is listening to person B?*	Listening	Speaker A is listening with attention to speaker B and follows the dialogue.
		Not Listening	Speaker A seems not to pay attention to what speaker B is saying.
		I don't know	It is unclear if speaker A is listening to speaker B or not.
Correctness	*Is the proposed response grammatically correct?*	Correct	The response does not contain any type of grammatical or structural error, any repetitions, misspellings or any other types of error.
		Not Correct	The response contains some grammatical or structural errors such as, repetitions, misspelling, any other types of error.
		I don't know	It is hard to identify if the response contains errors or not.

Table 4.1: The questions and possible answer options presented to the annotators for the evaluation of the response candidates in this version of the protocol.

Decisions For each criterion, the annotators are asked to select an answer from a 3-point Likert scale modeled as positive (eg. Correct, Appropriate), negative (eg. Not Correct, Not Appropriate), and *"I don't know"*. The purpose of the third choice, *"I don't know"*, is to avoid forcing non-deterministic and error-prone judgments on one of the other two options. That is, the non-expert annotator (in some cases nor the expert annotator) may not be able to make a deterministic decision due to the residual and inevitable ambiguity of the annotation task.

Explanations In order the obtain better insights into the capabilities and limitations of the models, we ask the annotators to explain their judgment by pointing out possible errors or rightness of a response. The explanation is asked for three of the criteria (listening is excluded) and mostly when the response is negatively evaluated or the annotator is not sure ("*I don't know.*"). In order to introduce the minimum amount of cognitive workload to the task, the annotators are asked to explain their judgment for each response right after evaluating a response candidate, through predefined options to select from, and/or free text. The list of predefined explanation options to select from and the cases for which the explanation is asked is presented in Table 4.2.

B) Annotation Characteristics

Another principal aspect of HE experiments is the annotation characteristics. Despite the importance of this aspect and its influence on the resulting quality, little attention is given to the careful design of the HE annotation task.

We can model the annotation task as the interactions of the human (in our setting

4.2. The HE Annotation Protocol

Quality Dimension	Annotators' Decision		Quality Sub-dimension
	Value	Explanation Options	
Appropriateness	Appropriate	❑ ***"The proposed response is coherent with the dialogue context."***	*Coherence*
		❑ Add free form text explanation	-
	Not Appropriate	❑ ***"The proposed response is not coherent with the dialogue context."***	*Incoherence*
		❑ Add free form text explanation	-
	I don't know	❑ **Please Add free form text explanation (required)**	-
Contextualization	Not Contextualized	❑ ***"The response is generic or does not contain any explicit or implicit reference to what it has been said in the dialogue context."***	*Genericness*
		❑ ***"The response is not consistent with the information contained in the dialogue context."***	*Hallucination*
		❑ Add free form text explanation	-
	I don't know	❑ **Please Add free form text explanation (required)**	-
Correctness	Not Correct	❑ ***"The response contains grammatical errors."***	*Grammaticality*
		❑ ***"The response contains one or more parts that are repetitive."***	*Repetition*
		❑ Add free form text explanation	-
	I don't know	❑ **Please Add free form text explanation (required)**	-

Table 4.2: The explanation options provided to the annotators to support their decisions. The annotators can select predefined option(s) and/or write a free-form text. Each explanation option refers to a sub-dimension that is used as an interpretation for the result analysis. The sub-dimensions are not presented to the annotators.

the annotator) with a task system (the evaluation). From the beginning of the task, the annotator tends to create a mental model of the task according to the properties and information she/he is presented to [45]. One of the main causes of issues in such settings is the gap between the user's and the designers' mental models [53, 89]. Furthermore, studies show high levels of cognitive workload in a task reduce the humans' ability to retrieve and exploit knowledge; meanwhile, reducing the mental workload helps to reduce the frequency of errors [34, 61, 93]. Therefore, it is necessary to carefully design the annotation task to ensure a controlled level of cognitive workload throughout the task and minimize the possibility of misunderstanding or ambiguity for the annotators by using well-explained guidelines, a simplified User Interface, and a clear annotation process.

Guidelines & Examples An important resource in crowd-sourcing annotation tasks is the guidelines, which have the objective to introduce the task to the annotator and instruct them about the process. The task guidelines and the examples must be written with a clear and simple structure in order to minimize possible ambiguities for the annotators and help them form a mental model in line with one of the task designers. The examples should be carefully selected to point out the possible ambiguities and difficulties during the annotation and to help the workers get familiar with the task. Our task guidelines include an introduction to the task, the definition and description of each criterion and corresponding answer sets, as well as examples of various scenarios and annotations.

User Interface We designed and implemented a User Interface (UI) for the task of HE, with the objective of an easy-to-use and intuitive platform that is extendable to other versions of the evaluation, presented in Figure 4.2. Throughout the task, a short version of the guidelines is always presented to the annotator with the possibility to

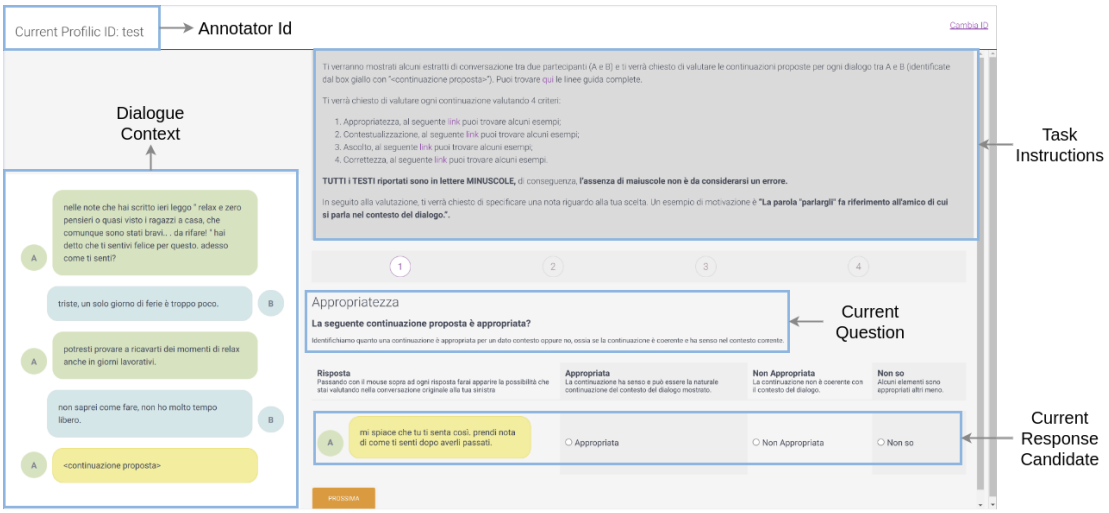

Figure 4.2: The user interface designed for the human evaluation of generated responses.

access the complete version via hyperlinks. During the evaluation, the corresponding dialogue context is shown to the annotator on the left, while the criterion question and the proposed response candidate are presented on the right, along with the name of the dimension, the definition of the dimension, and the possible decision values. In order to reduce the cognitive workload of the annotators, all candidates for a specific dialogue context are evaluated one by one for the same criterion after one another (i.e. the annotator evaluates all the candidates of the presented dialogue history for criterion A, and then all the same candidates regarding criterion B). In this way, the left side of the UI (dialogue history) remains unchanged so that the annotator does not have to go through the dialogue history several times, and focuses on each evaluation metric per sets of response candidates.

Internal Pilots Internal pilots can provide reliable feedback about the difficulty/-subjectivity of the task, the amount of time required to perform the task, and a threshold for the expected output quality of the task if done correctly. Internal pilots also help to detect and resolve possible ambiguities and issues in the task and its materials prior to the main task.

4.2.2 Annotator Recruitment

After designing the task, we need to recruit the required number of annotators to perform the task. In most cases, the annotation is done through crowd-sourcing. In that case, there are several aspects involved in the process of recruiting the crowd-workers that can affect the outcome quality including the sampling policy, the qualification, and the compensation.

Sampling In order to obtain reliable results, it is important to recruit the annotators from the correct target group. In the literature, selecting the annotators has been mostly conditioned by prerequisites such as location, language fluency, and level of education.

Qualification Karpinskaet al. [28] observed that when the annotators are sampled from workers in crowd-sourcing platforms, sampling conditions are not adequate as

they may be fulfilled inappropriately (for instance the use of VPNs to fake a certain location). Therefore, in addition to the mentioned prerequisites, it is essential to set up a qualification task for the workers. The qualification task helps the task designers to filter out contributors with low-quality performance and helps the crowd-workers to get familiar with the main task and the UI.

Compensation Proper compensation is an important extrinsic factor that can affect the performance of crowd workers, and the time it takes for the job to be selected and worked on by the workers [43, 61, 86]. Therefore, it is crucial to estimate properly and fairly the time and complexity needed to complete the task and set a fair wage in order to ensure proper compensation.

4.2.3 Task Execution

The execution of the main task is subject to continuous control of the progress and quality. In this phase, the agreement level among the annotators can indicate whether the outcome quality is maintained throughout the task. Sudden drops or jumps in the agreement level can be due to unbalanced difficulty among batches, or a low-quality contributor. While the former should be addressed using stratified sampling when designing the task, Riccardi et al. [66] observed that providing real-time feedback to the annotators helps them to recover their mistakes and improve their performance for the upcoming tasks.

4.2.4 Annotation Reporting

Howcroft et al. [22] highlight the lack of a standard for reporting the description and the results of HE experiments and point out the need for proper reporting of the evaluation details and results analysis. Furthermore, Ramirez et al. [61] stress the importance of reporting the crowd-sourcing experiment in a proper and standard way in order to facilitate the replicability of the experiment and the reproducibility of the results. We provide a checklist of aspects and elements that are necessary to be reported along with the final results in order to ensure a clear and transparent presentation of the protocol and possible outcomes. The characteristics of the task that should be reported are:

- Evaluation granularity (dialogue-level vs. response-level, comparative vs. absolute)
- Quality dimensions, their definitions, and corresponding questions
- Annotation format (item selection, free-form text, ranking, rating, etc.)

While the details regarding the recruitment of the crowd-workers include:

- Sampling criteria, the description of qualification task and acceptance\rejection criterion
- Number of workers recruited

Besides the mentioned details, there are certain statistics related to the execution of the evaluation task and its final outcome that should be reported to increase the credibility of the results. These statistics include:

- #Annotators participated in the study
- #Samples annotated in the study
- #Votes per each sample
- Inter-Annotator Agreement level & the metric used
- Workload allocated per annotator
- Demographic of the annotators
- Resource Utilization (time to perform the task, payment to the annotator, crowdsourcing platform)

4.3 Validation of the Protocol

We validate the proposed protocol by evaluating two Pre-trained Language Models (PLMs) for the Italian language for the task of response generation. The first model fine-tuned is iT5-Base [72], which has the same architecture as T5 PLM [60], pre-trained on a large Italian corpus. It consists of 12 layers per stack (encoder or decoder) with 220M parameters. The second model is GePpeTto [12] based on GPT-2 small [59], for the Italian language. The model consists of 12 layers of decoder and byte-pair encoding, with 117M parameters.

We fine-tune the two models using the dataset of follow-up (second-session) dialogues collected in Chapter 2. iT5-Base was fine-tuned using AdaFactor optimizer [80] and early stopping wait counter equal to 3, with batch size and dialogue history window equal to 4. GePpeTto was fine-tuned using AdamW optimizer [39] and early-stopping wait counter equal to 3, with batch size and dialogue history window equal to 2. 80% of the dataset was used as the fine-tuning training set, while 10% was used as the validation set for early stopping and parameter engineering, and the rest of the data, unseen 10%, was used as the test set (the splits were sampled at dialogue level to ensure no history overlap among splits). To evaluate the models in grounded response generation setting as well, we provide the user responses in the first dialogue sessions as unprocessed knowledge pieces for the response generation in the second session. The average length of knowledge with this representation is 126.7 tokens (this setting will be referred to in Section 5 as "*RAW*" representation of knowledge). We then fine-tuned the models for grounded generation via the same approach used by Zhao et al. [99]. The automatic evaluation of fine-tuned models is presented in Table 4.4.

4.3.1 Implementation

We implemented the proposed protocol to evaluate the performance of the two models via human crowd-workers.

Task Design We followed the Task Design step explained in subsection 4.2.1 closely. We then sampled 42 different dialogue histories from the fine-tuning test set (approximately 50%) for the evaluation (the length of histories varies from 2 to 4 turns) and sampled the responses of all models for each dialogue. We conducted two internal pilots using 5 dialogues (sampled from validation set) with 3 internal experts (the experts were not involved in the design of the task), as well as 3 internal non-expert annotators. After each pilot, feedback from both groups was collected and a few refinements were made to the UI and the guidelines.

Using the feedback obtained from the internal pilots regarding the difficulty of the task and the amount it takes to annotate the samples, we prepared the annotation batches so that each batch consists of approximately 10 dialogue histories of 4 turns in average, with 3 response candidates (including the ground truth) to evaluate for the next turn. During the internal pilots, each batch of 5 dialogues took an average of 15 minutes for the non-expert annotators. Therefore, we set the average required time to 35 minutes and the maximum time possible to annotate a batch to 90 minutes, in order to factor in the possible lower pace of non-expert annotators.

Recruiting Crowd-worker We used Prolific crowd-sourcing platform[2], and selected the crowd-workers using the following prerequisites:

- **Location**: Italy
- **Gender Distribution**: Available to All
- **First Language**: Italian
- **Minimum Approval rate**: 95%
- **Minimum complete submissions**: 20 jobs
- **Education**: Available to all
- **Expertise**: Available to all

In addition to the sampling policy, the annotators were asked to perform a qualification task. The task consisted of evaluating the response candidates for 5 dialogues (the same dialogues used in the internal pilots) in an identical setting to the main task. We considered the Inter-Annotator Agreement (IAA) of the internal non-expert annotators calculated by Fleiss' κ [16] as the threshold (0.21). In order to qualify each worker, we computed the agreement level between the internal annotators and the worker and if it was above the threshold, the worker was qualified for the main task.

Based on the workload and the estimated time required for the task, we set the wage as 4.67 pounds for 35 minutes, equal to 8 pounds per hour[3]. Qualified crowd-workers were also paid for the qualification task.

[2]Prolific: `https://www.prolific.co/`

[3]Prolific's Payment Principles mandates a fair and ethical payment to the workers with the minimum

Models	Inter-Annotator Agreement measured by Fleiss' κ				
	Appropriateness	*Contextualization*	*Correctness*	*Listening*	per Model
GePpeTto	0.27	0.14	0.64	0.15	0.32±0.10
+Knowledge	0.42	0.22	0.36	0.27	0.36±0.11
iT5-Base	0.24	0.19	0.06	0.18	0.27±0.04
+Knowledge	0.18	0.03	0.30	0.21	0.19±0.06
IAA per Dimension	0.30 ±0.10 **Fair**	0.15±0.05 **Poor**	0.41±0.20 **Moderate**	0.23±0.07 **Fair**	-

Table 4.3: The Inter-Annotator Agreement (IAA) level calculated by Fleiss' κ. The last row and last column represent the average IAA (and the standard deviation) per each of the criteria and each model, respectively. The low IAA on *Contextualization* indicates the high level of complexity and subjectivity in this criterion. In contrast, the moderate level of IAA is achieved over *Correctness* criterion, suggesting a lower level of subjectivity in the judgments.

4.3.2 Annotation Statistics

In total, 40 workers participated in the annotation task and 35 of them were qualified. The 42 samples to annotate were distributed in two batches of 11 and two batches of 10 samples. Each batch is annotated by 7 annotators and the annotators spent an average of 19 minutes for the qualification batch and 45 minutes for annotating the main batches. In addition to the decided compensations, one annotator was rewarded a bonus of two pounds since he/she informed us about an unexpected bug in the UI via email.

During the execution of the task, we calculated the agreement between each pair of annotators using Cohen's κ [9] as well as the agreement among all annotators in the same batch using Fleiss' κ [16] metrics. We further calculated the agreement among all annotators on strong judgments, by removing items that were labeled as "I don't know." by at least one annotator. Despite little fluctuations in the agreement level, no low-quality contributions were detected and the agreement level on different batches was consistent throughout the evaluation.

Table 4.3 presents the average Inter Annotator Agreement (IAA) measured by Fleiss' κ. The agreement is calculated per each model and criterion in each batch (for the 7 annotators who annotated the batch) and averaged over all batches. The results indicate that *Contextualization* and *Listening* are the two criteria with the highest levels of subjectivity and complexity. In contrast, high IAA over *Correctness* suggests that it has been easier for the annotators to assess the grammatical and structural aspects of the response samples.

4.3.3 Evaluation Results

Table4.4 presents the results of the HE based on the majority voting for each model. While the grounding generally improved the performance of iT5-Base, it worsened the performance of GePpeTto in all aspects. Nevertheless, it introduced grammatical and

of 6 pounds (8 dollars) per hour. While deploying the study on the platform, the task owner is prompted with recommended payment level for the study, for which our payment of 8 pounds per hour was labeled as "Good".

4.3. Validation of the Protocol

Models			Human Evaluation			
	nll	***ppl***	**Appropriateness**	**Contextualization**	**Correctness**	**Listening**
Ground Truth	-	-	100.0%	97.62%	97.62%	97.62%
GePpeTto	2.76	15.84	66.67%	69.05%	83.33%	64.29%
+Knowledge	2.79	16.33	59.52%	57.14%	83.33%	57.14%
iT5-Base	2.05	7.79	66.67%	73.81%	100.0%	66.67%
+Knowledge	2.04	7.70	80.95%	80.95%	85.71%	76.19%

Table 4.4: The automatic and human evaluation outcome of the fine-tuned models. The results are obtained by majority voting. The evaluations indicate that grounding mostly improves the performance of iT5 Base, while it worsens GePpeTto's performance. Note that the perplexity can not be compared among models since the pre-training data and thus the vocabulary distributions are not identical.

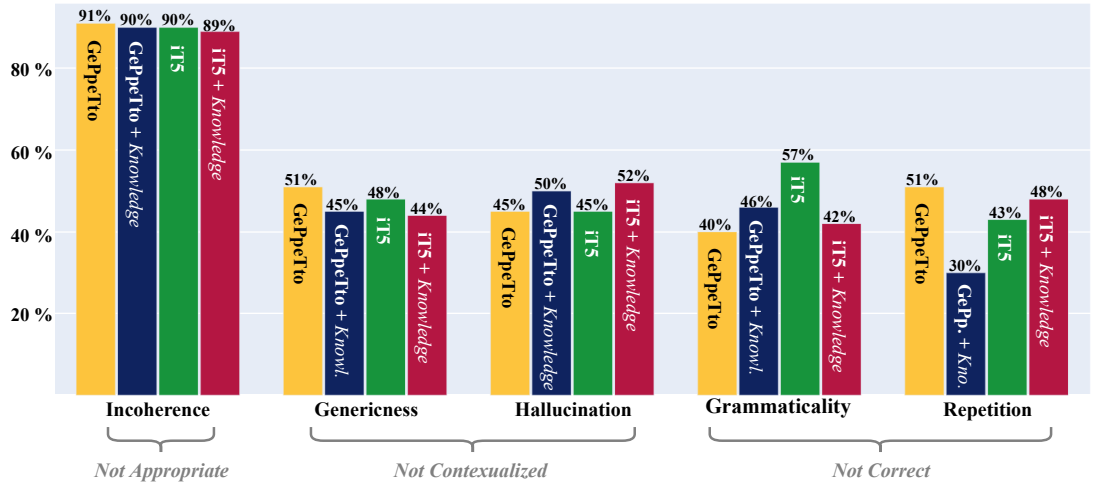

Figure 4.3: The sub-dimension errors selected by the annotators for the explanation of negative judgments in each criterion. Each bar represents the percentage of the times the error category (x-axis) was selected as the reason to reject the output of the corresponding model. The figure is obtained by considering all the votes (i.e. not majority voting). Note that the labels are not mutually exclusive.

structural errors in iT5-Base output. Moreover, grounding did not improve GePpeTto to generate more contextualized responses.

Figure 4.3 represents the sub-dimension errors that the annotators selected to explain their negative votes on the response candidates (The explanation option corresponding to each error is presented in Table 4.2). The figure is obtained by considering all the votes of the annotators on every response sampled from the models (each response is evaluated by 7 annotators, thus 294 votes in total). Therefore, for instance, while iT5-Base achieves 100% of "*Correctness*" by majority voting, there are 7 cases (out of 294) where the annotators labeled it as "Not Correct"; the selected reason in 4 cases was a grammatical error and in 3 cases a repetition in the response.

These results indicate that, regardless of the model, while grounding reduces the cases that a response is labeled as "*Not Contextualized*" due to being a *Generic* response, it increases the cases of *Hallucination* problem with almost the same propor-

tion. Nevertheless, the percentage of cases where a response is labeled as "*Not Appropriate*" due to being *Incoherent* is not affected by the grounding technique and all models suffer from this error equally. Furthermore, we observe that grounding slightly increases the cases in which a response by GePpeTto is labeled as "*Not Correct*" due to errors related to *Grammaticality*, while it considerably reduces the cases of *Repetition* in such responses.

In addition to the pre-defined explanations, in a few cases, the annotators also provided us with free-form explanations. Specifically, in 10% of the cases in which the model outputs were labeled as "*Not Correct*", the annotators provided us further explanations to indicate the exact grammatical error such as punctuation or subjunctive errors (Congiuntivo in Italian). In 5% of the times in which the model responses were considered "*Not Contextualized*" the annotators pointed out the exact part of the response which is mentioning a wrong event/participant or is in contradiction to the dialogue history. Lastly, in 10% of the cases where the response candidate was evaluated as "*Not Appropriate*" the annotators provided explanations to highlight the exact segment of the response that is not right or is ambiguous.

4.4 Conclusion

We presented a complete methodology for HE of generated responses to reach comparability and replicability of evaluation results among works and models. We unfolded the task of HE into four main executive steps and studied the necessary properties and actions to ensure a reliable evaluation while minimizing task complexity. We validated the protocol by evaluating two PLMs for the task of response generation with and without knowledge grounding, where we managed to identify the types and distributions of errors the models made. We publish the protocol and all its materials to the community and engage them to utilize, extend, and complement this protocol into further versions as a transparent protocol for HE of response generation models[4].

[4]https://github.com/sislab-unitn/Human-Evaluation-Protocol

CHAPTER 5

Response Generation in Longitudinal Dialogues

In this chapter of the thesis, we study the task of response generation in Longitudinal Dialogues (LD). LDs include the recollections of events, thoughts, and emotions specific to each individual in a sparse sequence of dialogue sessions. Thus, the model must generate a response specific to the individual user, that must be coherent with both the dialogue context and the previous dialogue sessions of the user.

We investigate the applicability of general-purpose Pretrained Language Models (PLM) for the task of response generation in LDs. We conversationally fine-tuned two recent PLMs, GePpeTto (GPT-2) [12] and iT5 [72], as a decoder-only and an encoder-decoder architecture, using the dataset of LDs collected in Chapter 2.

To improve the quality of machine responses, we experiment with the representations of the context in LDs for grounded response generation. We use the responses each individual user shared in the previous dialogue sessions with the system as personal knowledge, and evaluate whether grounding on such knowledge results in more appropriate and personal responses. In previously published research on grounded generation [25, 87, 99], the knowledge sequence is provided to the model as-is. In this work, we experiment with three different representations of the knowledge piece. The first two representations are a) *Raw* as unprocessed text, similar to the previously published research; and b) Bag of Head nouns (*BOH*) as a distilled syntactic representation of the knowledge. For the third representation, we use the Personal Space Graph (*PSG*) of the events and participants mentioned in the user responses, presented in Chapter 3. An example of a dialogue and different representations of the corresponding personal knowledge is shown in Figure 5.1.

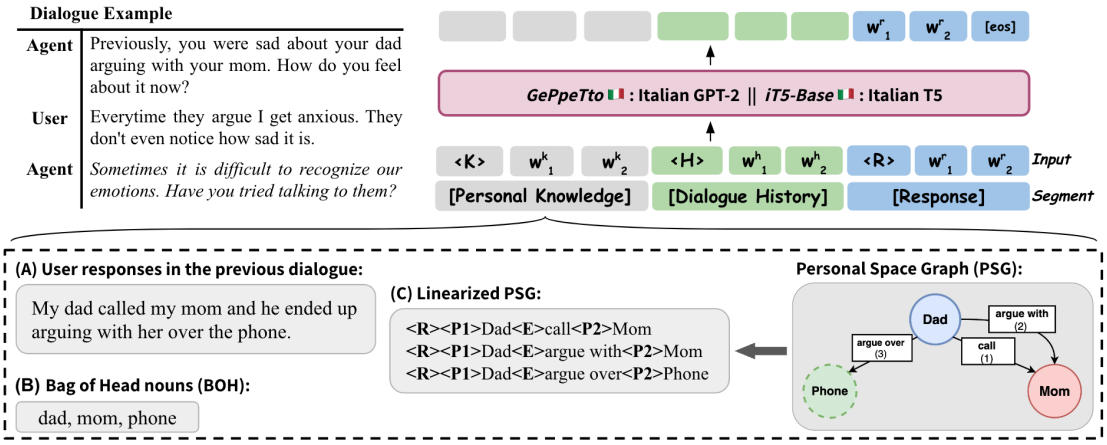

Figure 5.1: An example of a longitudinal dialogue. The user responses in the previous dialogue session are used as personal knowledge for grounded response generation. The knowledge is presented to the model as A) Unprocessed text (*RAW*); B) Bag of Head nouns (*BOH*); and C) Personal Space Graph (*PSG*) of events and their participants in linearized format. The model then encodes the dialogue history and the knowledge piece and generates a response candidate (the last agent turn in the dialogue example).

5.1 Background

Grounded Response Generation The application of PLMs has achieved acceptable performance in many tasks [82]. However, as end-to-end generative models, such models are known to suffer from the major issue of generating inappropriate and/or generic responses which can lead to ethical problems and low user engagement [97]. A recent research focus to address this problem is to condition the generation on both the dialogue history and a knowledge piece external to the dialogue history, to improve the generation quality [25, 99]. In this regard, Zhao et al. [99] proposed a model for grounded response generation using PLMs where fine-tuning the model and training of the knowledge selection module happens jointly. Similarly, Huang et al. [25] introduced a Transformer based model for open-domain dialogues with joint optimization of the knowledge selection module and response generation model. Hedayatnia et al. [21] proposed a Transformer-based model that initially generates an action plan consisting of the essential attributes for the response such as the set of dialogue acts and the most relevant knowledge sentence, and uses the plan to generate a grounded response. Rashkin et al. [62] proposed three criteria relevant to compliance of the generated response to the knowledge piece and studied the integration of these metrics into the model to increase controllability.

Personal Response Generation The research on personalized response generation in dialogues has been limited to persona descriptions and limited sets of user preferences and profiles. Zhang et al. [95] collected a dataset of open-domain dialogues using Amazon Mechanical Turk (AMT) workers conditioned on synthetic sets of 5 sentences as personas for each side of the dialogue. This dataset was used by Wolf et al. [87] and Kasahara et al. [29] to fine-tune GPT-2 architecture [59] for personal re-

sponse generation conditioned on the persona sentences. Modotto et al. [41] developed a model that utilizes meta-learning to learn the users' persona from the dialogues samples of the same user, rather than the persona descriptions. Huang et al. [24] developed a framework to adapt the attention weights on the input sequence dynamically to obtain a better representation of the persona.

While the mentioned work focused on personalization in open-domain dialogues, Joshi et al. [27] generated profiles consisting of gender, age, and food preference permutations for the user side in restaurant booking dialogues. This dataset was used by Siddique et al. [74] to fine-tune a PLM architecture to generate personalized responses in a task-based dialogue.

5.2 Experiments

5.2.1 Models

We fine-tuned two state-of-the-art PLMs using the dataset of LDs.

GePpeTto: Italian GPT-2 The first model we experimented with is GePpeTto [12], a PLM based on GPT-2 small (12 layers of decoder, 117M parameters) [59], trained for the Italian language (13 GB corpus size). We fine-tuned the model using AdamW optimizer [39] with an early-stopping wait counter equal to 3 and a history window of 2 last turns.

iT5: Italian T5 The second PLM in our experiments is iT5 [72], a PLM based on T5 [60], trained on the Italian portion of mC4 corpus (275 GB corpus size). We experimented with iT5-Small (12 layers, 60M parameters) and iT5-Base (24 layers, 220M parameters) [1]. We fine-tuned this model class using AdaFactor optimizer [80] with early stopping wait counter equal to 3 and a history window of 4 last turns.

5.2.2 Dataset

We fine-tuned the models using a dataset of LDs collected in Chapter 2. As mentioned, there are two dialogue sessions for each individual user in this dataset. In the first dialogue session, the system prompts the user to engage her in the recollection of daily life events the user has experienced. Throughout the interaction, the user shares details about the events and participants that have activated her emotions by answering a set of questions.

For each user, the first session is then followed by a follow-up dialogue. In the second dialogue session, the user tends to share more details about her feelings and the possible evolution of the previously mentioned events. Meanwhile, the listener provides personal suggestions and asks questions to expand or disambiguate previously stated facts or feelings. A mock-up example of a second dialogue session and the corresponding user response in the previous dialogue is shown in Figure 5.1. This dataset consists of 800 two-session dialogues with an average of 5 turns per dialogue.

[1] We were unable to use iT5-Large due to lack of GPU memory

5.2.3 Grounded Response Generation

We experimented with grounded response generation to improve the quality of the models' output. For each user, we extracted her responses in the first dialogue session as personal knowledge to ground the response generation for the second dialogue session. We experimented with three representations of the knowledge piece:

- **(A) RAW**: We provide the responses of the user in the previous dialogue as an unprocessed knowledge piece. The average length of knowledge with this representation is 126.7 tokens.
- **(B) Bag of Head nouns (BOH)**: We automatically parse the user responses with the spaCy[2] dependency parser and extract the head nouns as a distilled syntactic representation of the knowledge.
- **(C) Personal Space Graph (PSG)**: Using the approach proposed in Chapter 3, we represent the knowledge by the personal graph of the events and participants mentioned by the user. In this approach, the predicates in a sentence represent an event and its corresponding noun dependencies (subject, object) represent the participants. In this graph, the participants are the nodes while the predicates are the relations (edges) among the participants. We obtain a linear representation of the graph using an approach inspired by Ribeiro et al. [65]. The authors observe that providing a linearized representation of the graph to the PLMs results in outperforming the models with a graph-specific structural bias for the task of graph-to-text generation.

5.3 Evaluations

The fine-tuning of the models was done using NVIDIA GeForce RTX 3090 graphic card. The fine-tuning training set consisted of 80% of the dialogues (640 dialogues, 1284 samples with different turn levels), while the remaining data was split into 10% (80 dialogues, 160 samples with different turn levels) as the validation set for parameter engineering and early-stopping, and 10% as unseen test set. Each split was sampled at the dialogue level to guarantee no history overlap among splits. An example of a second dialogue session and the generated responses are presented in Appendix Table 5.5.

5.3.1 Automatic Evaluation

The results of the automatic evaluation of the models, presented in Table 5.1, show that incorporating the knowledge slightly increases the negative log-likelihood loss (*nll*) and consequently the perplexity scores of all models. The perplexity scores cannot be used to compare the performance between GePpeTto and iT-5 model classes as the vocabulary distributions in the pre-training phase of the two PLMs are not identical. However, the scores are comparable among iT5 variations as the same model class

[2]spaCy Library: spacy.io

Models	*nll*	*ppl*
GePpeTto	2.76	15.84
$+RAWKnowl.$	2.79	16.33
$+BOHKnowl.$	2.85	17.38
$+PSGKnowl.$	2.77	16.06
iT5-Small	2.18	8.84
$+RAWKnowl.$	2.19	8.95
$+BOHKnowl.$	2.18	8.88
$+PSGKnowl.$	2.19	8.93
iT5-Base	2.05	7.79
$+RAWKnowl.$	2.04	7.70
$+BOHKnowl.$	2.12	8.40
$+PSGKnowl.$	2.09	8.07

Table 5.1: Automatic evaluation of the models indicates that incorporating the knowledge slightly increases the models' perplexity (Perplexity scores can not be compared among models since the vocabulary distributions of pre-training data are not identical).

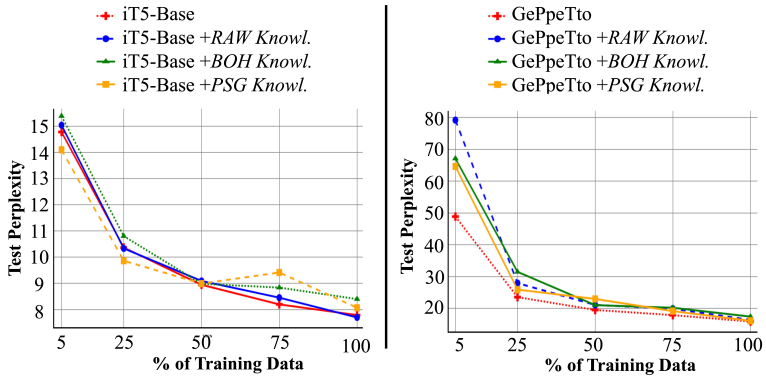

Figure 5.2: Perplexity score trends of the models over increasing size of the training set. The performance of GePpeTto variations is considerably improved after observing 50% of the fine-tuning training set.

pre-trained using the same data. In fact, the perplexity scores indicate that iT5-Base demonstrates a better performance than iT5-Small in all combinations with knowledge representations. Therefore, we select iT5-Base among the iT5 models and focus the rest of the analysis on GePpeTto and iT5-Base.

Considering the small size of the LD dataset compared to the data used in the pre-training phase, we studied the impact of fine-tuning the models by optimizing the models over increasing size of the training set. The extension of the training set was gradual (the small portions are subsets of the big portions) and the performance of models was evaluated by measuring the perplexity score on the unseen test set. The results are presented in Figure 5.2. The performance of both models is improved considerably after observing the first 25% and 50% of the train set, thus the fine-tuning has been more effective. However, in the second half of the data, both models show a steady trend while iT5-Base achieves a gradual improvement.

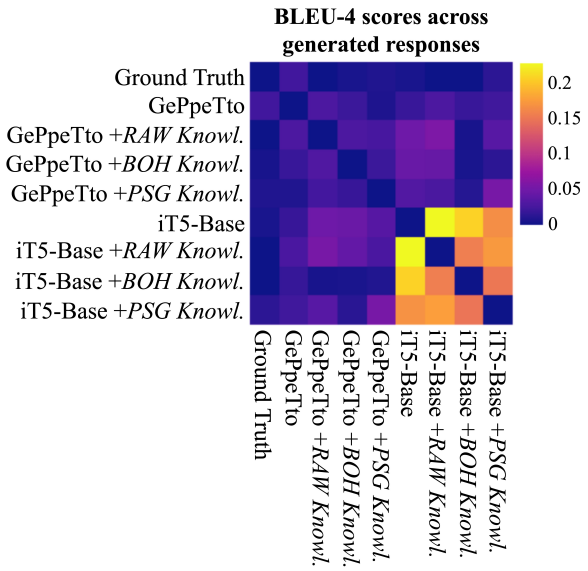

Figure 5.3: Lexical similarity among generated responses on the test set measured by BLEU-4 score. The results indicate a higher similarity among the responses generated by iT5-Base models.

To investigate the impact of grounding on the response lexicalization of the models, we measured the diversity in the generated responses for the test set samples via BLEU-4 score, Figure 5.3. We observed that there is a higher similarity among responses generated by iT5 models, while the responses generated by GePpeTto variations are more diverse. A similar finding has been observed in the literature about the performance of auto-regressive models compared to encoder-decoder architectures regarding novelty in sequence generation [4, 78]. Further, responses generated by iT5-Base with *BOH* and *PSG* representations have the lowest lexical similarity. The responses with the highest lexical similarity are generated by iT5-Base with no grounding and *RAW* representation. Nevertheless, there is a negligible lexical similarity between the generated responses and the ground truth.

5.3.2 Human Evaluation

We sampled 42 dialogue histories (approximately 50%) of the unseen test set and evaluated the generated responses via human judges using the protocol proposed in Chapter 4. We evaluated the responses according to four criteria:

- ***Correctness***: evaluating grammatical and syntactical structure of the response.
- ***Appropriateness***: evaluating the response to be a proper and coherent continuation with respect to the dialogue history.
- ***Contextualization***: evaluating whether the response refers to the context of the dialogue (not generic) or it consists of non-existing/contradicting information (hallucination cases).

Models	Inter Annotator Agreement Level measured by Fleiss' κ				
	Appropriateness	*Contextualization*	*Correctness*	*Listening*	IAA per Model
GePpeTto	0.27	0.14	0.64	0.15	0.32±0.10
+RAWKnowl.	0.42	0.22	0.36	0.27	0.36±0.11
+BOCKnowl.	0.23	0.05	0.31	0.11	0.27±0.05
+PSGKnowl.	0.30	0.39	0.34	0.26	0.42±0.06
iT5-Base	0.24	0.19	0.06	0.18	0.27±0.04
+RAWKnowl.	0.18	0.03	0.30	0.21	0.19±0.06
+BOCKnowl.	0.21	0.17	0.58	0.24	0.26±0.09
+PSGKnowl.	0.17	0.06	0.27	0.14	0.19±0.12
IAA per Dimension	0.31±0.09 **Fair**	0.20±0.06 **Poor**	0.43±0.20 **Moderate**	0.25±0.10 **Fair**	-

Table 5.2: Inter-Annotator Agreement (IAA) level calculated by Fleiss' κ. IAA is calculated per each model and criterion in each batch (4 batches, 7 annotators per batch) and averaged over all batches. Low IAA level for *Contextualization* suggests a high level of subjectivity in this criterion.

- ***Listening***: whether the generated response shows that the speaker is following the dialogue with attention.

The annotators were asked to evaluate the response candidates and select a decision for each criterion from a 3-point Likert scale as positive (eg. Correct, Appropriate), negative (eg. Not Correct, Not Appropriate), and "I don't know". We recruited 35 native Italian crowd-workers through Prolific crowd-sourcing platform[3]. The workers were asked to perform a qualification task consisting of evaluating 5 samples (sampled from the validation set) in an identical setting to the main task. For the main evaluation, each crowd-worker annotated 3 response candidates for 10 dialogue histories, and each sample was annotated by 7 crowd-workers. We also asked the annotators to motivate their decisions for appropriateness and contextualization criteria by providing an explanation to point out possible errors in the generated response. Moreover, the ground truth was also included in the candidate set to be evaluated.

The Inter Annotator Agreement (IAA) level measured by Fleiss' κ, presented in Appendix Table 5.2, indicates high levels of subjectivity and complexity in *Contextualization* criterion, suggesting that it has been difficult for the annotators to assess this aspect of the responses.

The results of the human evaluation of responses are presented in Table 5.3 (the scores are obtained by majority voting). The evaluation of GePpeTto models shows that grounding generally worsens the performance of GePpeTto, regardless of the representation format, as the best performance is achieved by GePpeTto with no knowledge grounding. Nevertheless, *BOH* and *PSG* representations slightly improve the grammatical correctness of this model. The highest level of *Contextualization* among grounded GePpeTto models is achieved by *PSG* representation. Regarding iT5-Base variations, the results indicate that grounding improves the models' performance considerably with respect to *Appropriateness*, *Contextualization*, and *Listening*. However,

[3]Prolific: `https://www.prolific.co/`

5.3. Evaluations

Models			Human Evaluation			
	nll	***ppl***	**Correctness**	**Appropriateness**	**Contextualization**	**Listening**
Ground Truth	-	-	97.62%	100.0%	97.62%	97.62%
GePpeTto	2.76	15.84	83.33%	**66.67%**	**69.05%**	**64.29%**
+RAW Knowl.	2.79	16.33	83.33%	59.52%	57.14%	57.14%
+BOH Knowl.	2.85	17.38	**92.86%**	45.24%	52.38%	42.86%
+PSG Knowl.	2.77	16.06	90.48%	54.76%	64.29%	50.00%
iT5-Base	2.05	7.79	**100.0%**	66.67%	73.81%	66.67%
+RAW Knowl.	2.04	7.70	85.71%	80.95%	80.95%	76.19%
+BOH Knowl.	2.12	8.40	92.86%	**80.95%**	85.71%	83.33%
+PSG Knowl.	2.09	8.07	95.24%	73.81%	**90.48%**	**83.33%**

Table 5.3: Human Evaluation of the fine-tuned models. The results indicate that grounding worsens the performance of GePpeTto, while it improves the performance of iT5-Base considerably. Moreover, *PSG* representation of knowledge achieves the best performance among grounded iT5-Base variations.

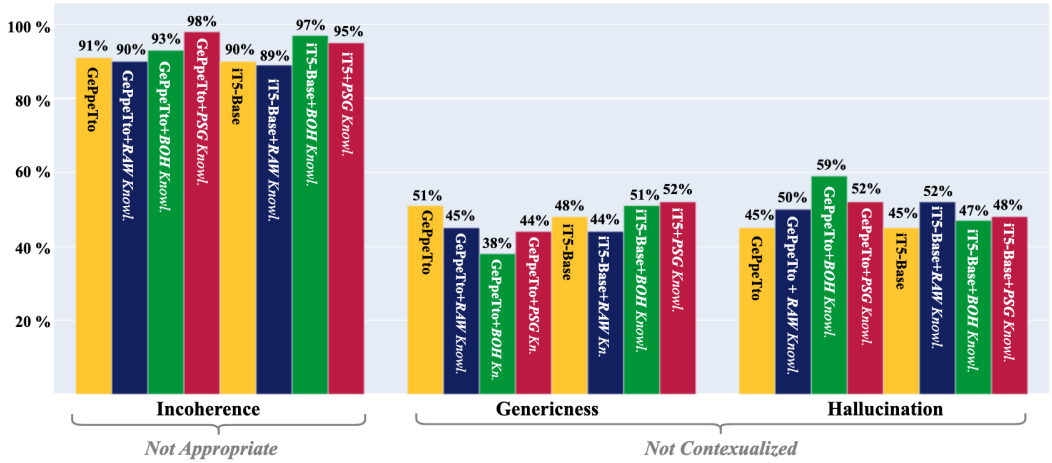

Figure 5.4: Explanations selected by the crowd-workers to motivate their negative judgments in *Appropriateness* and *Contextualization* criteria, represented by the percentage of the times the error category (x-axis) was selected. The figure is obtained by considering all the votes (i.e. not majority voting). Note that the labels are not mutually exclusive.

it decreases the model's *Correctness* with the highest decrease caused by *RAW* representation. *PSG* representation achieves the highest level of *Contextualization* and *Listening* overall, besides the highest level of *Correctness* among grounded models. Therefore, refined representations of the knowledge (*BOH* and *PSG*) generally result in better performances compared to *RAW* representation. Nevertheless, there is still a huge gap between the performance of the best-performing model and the ground truth, suggesting the grounded PLMs are not suitable dialogue models for LDs in the mental health domain.

To gain better insight into the errors made by each model, we investigated the reasons provided by the annotators for their judgments. These results, presented in Figure 5.4, are complementary to the evaluation decisions, Table 5.3, and point out the errors

Models	*Knowl.*	*History*
iT5-Base		
$+RAWKnowl.$	44.6%	55.4%
$+BOHKnowl.$	39.5%	60.5%
$+PSGKnowl.$	38.7%	61.3%

Table 5.4: Percentage of tokens with significant contribution to the generation (top-25%) in each segment of the input vector for each model.

that resulted in the negative evaluation of a response by the annotators. The analysis shows that grounding reduces the cases of genericness in rejected responses by GePpeTto while it slightly escalates this issue in iT5-Base rejected responses. Moreover, the rejected responses of iT5-Base with *RAW* representation were more hallucinated than other representations. Nevertheless, grounding does have any positive impact on the cases of incoherence in rejected responses of the PLMs.

5.3.3 Generation Explainability

According to the human evaluation results, iT5-Base with knowledge grounding achieves the best performance among PLMs. We investigated the contribution of personal knowledge and different representations on the performance of the model at inference time. We studied the attribution scores of the input tokens using the Integrated Gradients technique [73, 77] based on backward gradient analysis. We experimented with two thresholds for the attribution scores:

- **Positive Contribution**: Based on the assumption that elements with positive scores have a positive influence on the model's performance, we investigated the tokens with positive attribution scores, However, tokens with small attribution scores have negligible contributions and thus this analysis can be noisy.
- **Significant Contribution**: To identify the tokens with significant contributions to the generation, we selected the top-25% of the tokens in the input sequence (knowledge and history) according to their attribution score. We then investigated what portion of these tokens belong to each segment of the input vector. For a fair comparison, the values are normalized over the segment length.

According to Positive Contribution analysis, 74% of the tokens in the *RAW* representation have a positive contribution to the generation with the majority (30%) of tokens being verbs and nouns. This percentage for *BOH* (Bag of Head Nouns) representation changes to 79.0%. This result suggests the importance of nouns for the model inference. Regarding the *PSG* representation, 55.6% of the tokens have a positive contribution to the generation (excluding the tags used for linearization), with the majority (68%) of tokens being events rather than participants.

The analysis of the tokens with significant contributions is presented in Table 5.4. Regarding the model with *RAW* representation, the percentage of tokens with high attribution scores is almost balanced between the knowledge and history segments. How-

ever, for the models with refined representations of knowledge (*BOH* and *PSG*), the dialogue history contains moderately more significantly contributing tokens.

5.4 Conclusion

We studied the task of response generation in Longitudinal Dialogues (LD), where the model should learn about the user's thoughts and emotions from the previous dialogue sessions and generate a personal response that is coherent with respect to the user profile and state, the dialogue context, as well as the previous dialogue sessions. We fine-tuned two state-of-the-art PLMs for Italian, using a dataset of LDs in the mental health domain. We experimented with grounded generation using user responses in the previous dialogue session as user-specific knowledge. We investigated the impact of different representations of the knowledge, including a graph representation of personal life events and participants mentioned previously by the user.

Our evaluations showed there is still a huge gap between the performance of the general-purpose PLMs with knowledge grounding and the ground truth. Nevertheless, we observed that a) refined representations of the knowledge (such as *BOH* and *PSG*) can be more informative and less noisy for a grounded generation; b) the encoder-decoder model exhibited more diversity in the outputs compared to the auto-regressive model; c) knowledge grounding reduces the cases of genericness in response, though it can result in more hallucinated responses.

Dialogue Example		
Knowledge	**User Responses from Previous Dialogue:**	I only worked in the morning, after lunch I went to the bank with Martina to activate her prepaid card since the one she is using now is about to expire. I accompanied Paolo to fix the bike and then I went to the dentist. When I came back, I made sauce for the week and then I ironed. I ran a lot but it was not a very stressful day. I hope to be able to go to bed early to be able to rest!"
Dialogue History	**Dialogue Stimulus:**	I remember you had a positive feeling about what I read in your notes [yesterday] "*I only worked in the morning, after lunch I went to the bank with Martina to activate her prepaid card since the one she is using now is about to expire. I ran a lot but it was not a very stressful day. I hope to be able to go to bed early to be able to rest!*". Do you have any new thoughts on this?"
	User:	I think that working in the morning and in the afternoon was not tiring, actually it was pleasant. I was also able to go to bed early enough, and I am well rested.
Response Candidates	**Ground Truth**	Good! Did you even manage to spend time with your daughter?
	GePpeTto	I see. Do you often find yourself away from each other?
	+RAW Knowl.	I see. Maybe you could try doing more flexible jobs to increase your anxiety.
	+BOH Knowl.	I'm glad to hear that! Have a nice day!
	+PSG Knowl.	Good, it's important to know yourself.
	iT5-Base	I'm glad you feel rested enough. Have you tried any relaxation exercises?
	+RAW Knowl.	I'm glad you were able to go to bed early. If you want to write me about this, we can talk about it anytime.
	+BOH Knowl.	I'm glad you rested well. Did you try any relaxation exercises?
	+PSG Knowl.	I'm happy to you rested well. If you want to write more about this, we can talk about it anytime.

Table 5.5: Example of a second dialogue session, the corresponding user turns in the first session as personal knowledge, and the generated responses (English translation).

CHAPTER 6

Conclusions

In this thesis, we studied the design and training of dialogue models for Longitudinal Dialogues (LD).

Our first contribution was discussed in Chapter 2, where we presented a dialogue data acquisition methodology for LDs unique to the individual user, to address the problem of data scarcity for personal multi-session conversations. We collected a data set of LDs consisting of two dialogue sessions for each individual user and investigated the appropriateness of the collected dialogues for developing dialogue systems to hold LDs. This dataset was further used for developing a personal healthcare agent which was deployed in two clinical pilot studies, including the first registered Randomized Control Trial using a dialogue system application.

In Chapter 3, we presented an unsupervised approach to address the need for constructing user models specific to each individual user to carry out LDs. The developed model extracts the user's real-life events and participants from her responses throughout each interaction and presents them as the user's Personal Space Graph. We evaluated the performance of the model for the Italian and English languages using a dataset of personal narratives and a dataset of common sense stories. Afterward, to obtain a more informative and concise user model via her life events and participants, we studied the novelty of the events presented in user responses. We proposed a novel task of new event detection as they unfold in a narrative, by annotating a dataset of personal narratives with new events at the sentence level, and developing neural and non-neural baselines for the task of new event detection.

In Chapter 4, we addressed the incomparability and ambiguity of Human Evaluation tasks in the literature, by presenting a complete protocol for transparent and replicable evaluation of response generation models using human judges. We unfolded the

evaluation task into four executive steps as 1) Task Design; 2) Annotator Recruiting; 3) Task Execution; and 4) Annotation Reporting. We investigated the crucial aspects at each step to maximize the reliability and replicability of the evaluation while minimizing the task difficulty and complexity. We validated the protocol by evaluating two pre-trained language models for the task of response generation with and without knowledge grounding, and identified the types and distributions of the errors the models made.

Last, in Chapter 5 we investigated whether general-purpose pre-trained language models are appropriate for response generation in LDs. We fine-tuned two models, GePpeTto (Italian GPT-2) and iT5 (Italian T5) using the collected dataset of LDs. We experimented with grounded response generation using different representations of the personal knowledge extracted from previous dialogue sessions of the user, including Personal Space Graph representation. We evaluated the models using automatic and human evaluations and studied the contribution of knowledge in response generation via explainability studies.

6.1 Limitations & Future Directions

The reproducibility of the annotation tasks in Chapters 2 and 3 may be subject to variability due to the fact that the task was done by internal annotators and not through crowd-sourcing techniques. Moreover, the presented annotation methodologies may need to be refined for other languages as the dialogue data collection was done in Italian language and the new event annotation was done in English.

There might be language-specific limitations in the performance of the models in Chapter 5. Furthermore, GePpeTto (based on GPT-2 small) is the only candidate for auto-regressive models for the Italian language at the time of this thesis. Therefore, the performance of the model may be limited due to the small number of parameters. While we were unable to experiment with the iT5-Large model due to computation power limitations, iT5 and GePpeTto are the only generative pre-trained language models available for the Italian language at the time of this thesis.

We believe an interesting future work would be to collect a larger dataset of LDs including several dialogue sessions with users in different age groups. Regarding the Personal Space Graph model, we plan to annotate a subset of the dataset and evaluate the quality of the graph obtained using the unsupervised model. Further, we plan to improve the performance of the Personal Space Graph model by 1) modeling the temporal information of the events in the user response; 2) including the classification of new events in the pipeline, to obtain a more concise representation of the user model.

Bibliography

[1] Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. Keep me updated! memory management in long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[3] Anya Belz, Simon Mille, and David M. Howcroft. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *INLG*, 2020.

[4] Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroglu, and Marco Guerini. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049. Association for Computational Linguistics, December 2022.

[5] Nathanael Chambers and Dan Jurafsky. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[6] Nathanael Chambers and Dan Jurafsky. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[7] José Coch. Evaluating and comparing three text-production techniques. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996.

Bibliography

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

[9] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46, 1960.

[10] Morena Danieli, Tommaso Ciulli, Seyed Mahed Mousavi, Giuseppe Riccardi, et al. A conversational artificial intelligence agent for a mental health care app: Evaluation study of its participatory design. *JMIR Formative Research*, 5(12):e30053, 2021.

[11] Morena Danieli, Tommaso Ciulli, Seyed Mahed Mousavi, Giorgia Silvestri, Simone Barbato, Lorenzo Di Natale, Giuseppe Riccardi, et al. Assessing the impact of conversational artificial intelligence in the treatment of stress and anxiety in aging adults: Randomized controlled trial. *JMIR Mental Health*, 9(9):e38067, 2022.

[12] Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. Geppetto carves italian into a language model. *arXiv preprint arXiv:2004.14253*, 2020.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[14] Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online, July 2020. Association for Computational Linguistics.

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

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

[17] Chris Fournier and Diana Inkpen. Segmentation similarity and agreement. *arXiv preprint arXiv:1204.2847*, 2012.

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

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

[20] Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Aiswarya Baiju, Bing Liu, Ben Gerber, Lisa Sharp, Nadia Nabulsi, and Mary Smart. 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, 2020.

[21] Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 412–421, Dublin, Ireland, December 2020. Association for Computational Linguistics.

[22] David M. Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *INLG*, 2020.

[23] Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. Grade: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, 2020.

[24] Qiushi Huang, Yu Zhang, Tom Ko, Xubo Liu, Bo Wu, Wenwu Wang, and Lilian Tang. Personalized dialogue generation with persona-adaptive attention. *arXiv preprint arXiv:2210.15088*, 2022.

[25] Xinxian Huang, Huang He, Siqi Bao, Fan Wang, Hua Wu, and Haifeng Wang. PLATO-KAG: Unsupervised knowledge-grounded conversation via joint modeling. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 143–154, Online, November 2021. Association for Computational Linguistics.

[26] Karen Sparck Jones and Julia R Galliers. Evaluating natural language processing systems: An analysis and review. 1995.

[27] Chaitanya K Joshi, Fei Mi, and Boi Faltings. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*, 2017.

[28] Marzena Karpinska, Nader Akoury, and Mohit Iyyer. The perils of using mechanical turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, 2021.

[29] Tomohito Kasahara, Daisuke Kawahara, Nguyen Tung, Shengzhe Li, Kenta Shinzato, and Toshinori Sato. Building a personalized dialogue system with prompt-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 96–105, Hybrid: Seattle, Washington + Online, July 2022. Association for Computational Linguistics.

[30] Evgeny Kim and Roman Klinger. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[31] Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

[32] David B Kronenfeld. Scripts, plans, goals, and understanding: an inquiry into human knowledge structures by roger c. schank and robert p. abelson. *Language*, 54(3):779–779, 1978.

[33] James Lester and Bruce Porter. Developing and empirically evaluating robust explanation generators: The knight experiments. *Computational Linguistics*, 23(1):65–101, 1997.

[34] Nancy G Leveson. *Engineering a safer world: Systems thinking applied to safety*. The MIT Press, 2016.

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

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

[37] Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, 2016.

[38] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. 2019.

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[40] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 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, Prague, Czech Republic, September 2015. Association for Computational Linguistics.

[41] Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459, Florence, Italy, July 2019. Association for Computational Linguistics.

[42] William C Mann and Sandra A Thompson. *Discourse description: Diverse linguistic analyses of a fund-raising text*, volume 16. John Benjamins Publishing, 1992.

[43] Winter Mason and Siddharth Suri. Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1):1–23, 2012.

[44] Shikib Mehri and Maxine Eskenazi. Usr: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, 2020.

[45] Neville Moray. Identifying mental models of complex human–machine systems. *International Journal of Industrial Ergonomics*, 22(4-5):293–297, 1998.

[46] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for

deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California, June 2016. Association for Computational Linguistics.

[47] Alexander P. D. Mourelatos. Events, processes, and states by alexander p. d. mourelatos. *Linguistics and Philosophy*, 2(3):415–434, 1978.

[48] Seyed Mahed Mousavi, Alessandra Cervone, Morena Danieli, and Giuseppe Riccardi. Would you like to tell me more? generating a corpus of psychotherapy dialogues. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 1–9, Online, June 2021. Association for Computational Linguistics.

[49] Seyed Mahed Mousavi, Roberto Negro, and Giuseppe Riccardi. An unsupervised approach to extract life-events from personal narratives in the mental health domain. In *CLiC-it*, 2021.

[50] Seyed Mahed Mousavi, Gabriel Roccabruna, Michela Lorandi, Simone Caldarella, and Giuseppe Riccardi. Evaluation of response generation models: Shouldn't it be shareable and replicable? In *Proceedings of the Second Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2022)*. Association for Computational Linguistics, 2022.

[51] Seyed Mahed Mousavi, Shohei Tanaka, Gabriel Roccabruna, Koichiro Yoshino, Satoshi Nakamura, and Giuseppe Riccardi. Whats new? identifying the unfolding of new events in narratives. *arXiv preprint arXiv:2302.07748*, 2023.

[52] Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman. An annotation scheme for information status in dialogue. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA).

[53] Donald A Norman. *The psychology of everyday things.* Basic books, 1988.

[54] Horea-Radu Oltean, Philip Hyland, Frédérique Vallières, and Daniel Ovidiu David. 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, 2017.

[55] Desmond C. Ong, Zhengxuan Wu, Zhi-Xuan Tan, Marianne Reddan, Isabella Kahhale, Alison Mattek, and Jamil Zaki. Modeling emotion in complex stories: The stanford emotional narratives dataset. *IEEE Transactions on Affective Computing*, 12(3):579–594, 2021.

[56] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

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

[58] Ellen F. Prince. The zpg letter: Subjects, definiteness, and information-status. pages 295–325. John Benjamins, 1988.

[59] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[60] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[61] Jorge Ramírez, Burcu Sayin, Marcos Baez, Fabio Casati, Luca Cernuzzi, Boualem Benatallah, and Gianluca Demartini. On the state of reporting in crowdsourcing experiments and a checklist to aid current practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–34, 2021.

[62] Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718, Online, August 2021. Association for Computational Linguistics.

[63] Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[64] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 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, 2019.

[65] Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. Investigating pretrained language models for graph-to-text generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 211–227, Online, November 2021. Association for Computational Linguistics.

[66] Giuseppe Riccardi, Arindam Ghosh, SA Chowdhury, and Ali Orkan Bayer. Motivational feedback in crowdsourcing: a case study in speech transcription. In *INTERSPEECH*, pages 1111–1115, 2013.

[67] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

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

[69] Lorenza Russo, Sharid Loáiciga, and Asheesh Gulati. Improving machine translation of null subjects in italian and spanish. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 81–89, 2012.

[70] Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39, 2022.

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

[72] Gabriele Sarti and Malvina Nissim. It5: Large-scale text-to-text pretraining for italian language understanding and generation. *arXiv preprint arXiv:2203.03759*, 2022.

[73] Gabriele Sarti, Ludwig Sickert, Nils Feldhus, and Oskar van der Wal. Inseq: An interpretability toolkit for sequence generation models, January 2023.

[74] AB Siddique, MH Maqbool, Kshitija Taywade, and Hassan Foroosh. Personalizing task-oriented dialog systems via zero-shot generalizable reward function. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1787–1797, 2022.

[75] Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97, 2022.

[76] Josef Steinberger, Karel Jezek, et al. Using latent semantic analysis in text summarization and summary evaluation. *Proc. ISIM*, 4(93-100):8, 2004.

[77] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 06–11 Aug 2017.

[78] Serra Sinem Tekiroğlu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[79] David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. Adapting document-grounded dialog systems to spoken conversations using data augmentation and a noisy channel model. *arXiv preprint arXiv:2112.08844*, 2021.

[80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[81] Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. Paradise: A framework for evaluating spoken dialogue agents. *arXiv preprint cmp-lg/9704004*, 1997.

[82] Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. Pre-trained language models and their applications. *Engineering*, 2022.

[83] Yanmeng Wang, Wenge Rong, Jianfei Zhang, Yuanxin Ouyang, and Zhang Xiong. Knowledge grounded pre-trained model for dialogue response generation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[84] Charles Welch, Verónica Pérez-Rosas, Jonathan K Kummerfeld, and Rada Mihalcea. Learning from personal longitudinal dialog data. *IEEE Intelligent systems*, 34(4):16–23, 2019.

[85] Charles Welch, Verónica Pérez-Rosas, Jonathan K Kummerfeld, and Rada Mihalcea. Look who's talking: Inferring speaker attributes from personal longitudinal dialog. *arXiv preprint arXiv:1904.11610*, 2019.

[86] Mark E Whiting, Grant Hugh, and Michael S Bernstein. Fair work: Crowd work minimum wage with one line of code. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 197–206, 2019.

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

[88] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[89] Bingjun Xie, Jia Zhou, and Huilin Wang. How influential are mental models on interaction performance? exploring the gap between users' and designers' mental models through a new quantitative method. *Advances in Human-Computer Interaction*, 2017, 2017.

[90] Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[91] Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[92] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095, 2007.

[93] Marco A Zenati, Lauren Kennedy-Metz, and Roger D Dias. Cognitive engineering to improve patient safety and outcomes in cardiothoracic surgery. In *Seminars in thoracic and cardiovascular surgery*, volume 32, pages 1–7. Elsevier, 2020.

[94] Yutao Zeng, Xiaolong Jin, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. Event coreference resolution with their paraphrases and argument-aware embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3084–3094, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[95] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 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, 2018.

[96] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[97] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online, July 2020. Association for Computational Linguistics.

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

[99] Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online, November 2020. Association for Computational Linguistics.

[100] Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. 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 3363–3369, Hong Kong, China, November 2019. Association for Computational Linguistics.