Automatic Evaluation of Non-task Oriented Dialog Systems by Using Sentence Embeddings Projections and Their Dynamics

Mario Rodríguez-Cantelar, Luis Fernando D'Haro, and Fernando Matía

Abstract Human-Machine interaction through open-domain conversational agents has considerably grown in the last years. These social conversational agents try to solve the hard task of maintaining a meaningful, engaging and long-term conversation with human users by selecting or generating the most contextually appropriated response to a human prompt. Unfortunately, there is not a well-defined criteria or automatic metric that can be used to evaluate the best answer to provide. The traditional approach is to ask humans to evaluate each turn or the whole dialog according to a given dimension (e.g. naturalness, originality, appropriateness, syntax, engagingness, etc.). In this paper, we present our initial efforts on proposing an explainable metric by using sentence embedding projections and measuring different distances between the human-chatbot, human-human, and chatbot-chatbot turns on two different sets of dialogues. Our preliminary results show insights to visually and intuitively distinguish between good and bad dialogues.

1 Introduction

A social conversational agent is an automatic program designed to talk with human users about social or open topics (chitchat). In order to fulfill its work, this system must perform contextual modeling of syntactic, semantic and pragmatic information provided by the user along the different turns and answer accordingly with sentences that could maintain the coherence, naturalness, engagingness, humanness,

---

M. Rodríguez-Cantelar (✉) · F. Matía
Centre for Automation and Robotics (CAR) UPM-CSIC - Intelligent Control Group (ICG), Universidad Politécnica de Madrid, Madrid, Spain
e-mail: mario.rcantelar@alumnos.upm.es

F. Matía
e-mail: fernando.matia@upm.es

L. F. D'Haro
Information Processing and Telecommunications Center (IPTC) - Speech Technology Group, Universidad Politécnica de Madrid, Madrid, Spain
e-mail: luisfernando.dharo@upm.es

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
L. F. D'Haro et al. (eds.), *Conversational Dialogue Systems for the Next Decade*, Lecture Notes in Electrical Engineering 704, https://doi.org/10.1007/978-981-15-8395-7_6

and expectations of the users. In recent years, we have seen an exponential growth of research for chatbots to provide effective solutions to accomplish domain-specific tasks (e.g. buying movie tickets, play music or TV shows, recommend items, etc.), as well on domain-independent tasks (i.e. chitchat) where the incorporation of persona, emotion, and knowledge-based profiles is an active open research area to produce social-oriented chatbots.

Unfortunately, research in this area is highly limited due to multiple factors such as scarce number of training resources, intrinsic difficulties for modeling the human language, and the lack of automatic metrics that can model several dimensions, i.e. not only well-formed (syntactic) or correct (semantic) answers, but that can also provide explainability capabilities especially for non-task oriented chatbots. Traditionally, conversational systems are evaluated by means of subjective evaluations done by multiple users. However, this process is tedious, costly and slow, making difficult the faster development of current ML-based dialogue systems. Current objective metrics imported from related areas like machine translation or summarizing are being used [1], such as BLEU, ROUGE, o CIDEr, which calculates different distances between the sentence embeddings for the hypothesis reference and ground-truth answers, or the chatbot answer and the human prompt (e.g. RUBER [2]). Sadly, these metrics do not correlate well with human evaluations [3] making necessary to carry out a deeper analysis of the evaluation process itself and propose new ones.

In this paper, we will continue our previous work on evaluating generative conversational systems. In [4], we implemented and compared different DNN-based chatbots trained with different datasets and evaluated, on different dimensions and on a turn-by-turn basis, by several users through a subjective survey. In [5, 6], we proposed an objective metric for evaluating dialogue systems based on linearly measuring the fluency (syntax) and adequacy (semantic) of the generated responses and their similarity with given ground-truth references. In this paper, we are moving into a new approach where contextual information and the dynamics of the dialogue are considered. Although our results are preliminary, we observed interesting patterns and correlations that could provide new insights to develop a new metric. Our metric is inspired by [7] where two systems are evaluated at the end of the dialogue and in comparison, with another one. In our case, we propose three evaluations: (a) the Pearson correlation between human evaluations and Euclidean distances between the prompt and response turn pairs, (b) by comparing the accumulated Euclidean distances between the sentence embeddings for the same agent along all turns (i.e. evolution trace), and (c) the accumulated Euclidean distances for the pairs of prompt-response turn, and the next response-prompt turn (coherence). The study shows comparative results for these metrics between a human-chatbot interaction and human-human dialogues.

This paper is organized as follows. In Sect. 2, we describe the datasets, chatbot, and human evaluation used in this study. Then, in Sect. 3 we explain the mechanisms used for generating the sentence embeddings, projections, and metrics. In Sect. 4, we show our experiments, results, and analysis. Finally, in Sect. 5 the conclusions and future work.

2 Related Work: Datasets, Chatbot and Human Evaluation

For this project, we developed a generative-based chatbot [4] trained on Open Subtitles dataset [8] using a Seq2Seq [9] approach with bidirectional GRUs [10] and with Attention [11] and Beam Search [12] mechanisms to improve the quality of the responses. The final model consisted of 4 hidden layers with 256 hidden units, a 100K vocabulary, max sentence length set to 50 words, and adaptive learning rate.

To evaluate the quality of the responses, we carried out a subjective evaluation with a total of 25 evaluators with ages ranging between 18–35 years old. These evaluators were asked to read the dialogue shown in Table 2 and then for each chatbot's answer to evaluate four dimensions or aspects using a binary scale, i.e. assigning a 1 when they agreed and 0 otherwise. The reason for selecting this binary scale was to reduce the annotation effort. In detail, the four dimensions were:

Semantic	Meaning that the chatbot's answer is appropriated given the dialogue context and last user's prompt.
Syntactic	The chatbot's answer is grammatically correct.
Correctness	The chatbot's answer is not just topically adequate w.r.t the user's prompt, but also right. E.g. if asked for 1 + 1 the system not only answer with a number, but with number 2.
Specificity	The chatbot's answer is specific to the user's prompt, not a generic, neutral or safe answer.

Then, the mean and standard deviation for the different dimensions were calculated, together with the general punctuation as the sum of the four. Our results showed that the chatbot presented a high Semantic (81.29%) and Syntactic (86.78%) results and lower Correctness (70.10%) and Specificity (76.88%), similar to the results reported in [9]. Additionally, we calculated the total score as the sum of the unbiased four scores and then calculated the mean score for all of the evaluators (i.e. the global avg. chatbot's score was 3.15 in a scale from 0 to 4).

On the other hand, as a contrastive dataset, we will use a subset of dialogues (i.e. 50 randomly selected dialogues) from the Persona-Chat dataset [13] which consists of 162K utterances over 11K dialogues, and where around 1.1K persona profiles were defined to generate human-human introduction dialogues where the two participants shared likes and some background information. Human evaluations done during the data collection showed an avg. of 4.3 for fluency, 4.25 for engagingness, and 4.4 for consistency over a 5 points scale, which can be considered very good.

3 Embedding Projections and Proposed Metrics

To generate sentence embeddings for each turn in the dialogue, we used the ConveRT dual-encoder model [14] based on Transformers [15] due to its excellent reported results, reduced model size, and efficiency. This model uses sub-word units to reduce problems with OOVs, a set of Transformer blocks for the encoder and it has been optimized to consider the context during the projections for the down-stream task of segment prediction. The model was pre-trained using the Reddit conversational corpus [16] and fine-tuned for the DSTC7 answer classification task on the Ubuntu dialogue corpus [17]. The advantage of these sentence embeddings is that they encapsulate low-level (syntactic) and high-level (semantic) information from the words used in the sentence and the dialogue history. In our study, the estimated sentence embeddings had 512 dimensions that were standardized to have zero mean and unit variance. Since ConveRT has been trained on different dialogue datasets, it has shown better-reported results across different applications in comparison with other encoders such as BERT [18] or USE [19]; besides, this model is wrapped using a convenient interface that allows the encoding of sentences by considering also contexts and responses [20].

Prompt-Answer Correlation: In our study, we first calculated the Pearson correlation between the unbiased averaged human evaluation total score (S) assigned to each chatbot's answer and the Euclidean distance between the human's prompt and chatbot's answer sentence embeddings for each turn pair (P) in the dialogue shown in Table 2. In concrete we used Eq. 1:

$$Pearson\ Correlation\ (P, S) = corr\ (\mathbf{dist}\ (\mathbf{p},\ \mathbf{r}),\ \mathbf{AvgScore}) \tag{1}$$

Where p and r are the human's prompts and chatbot's responses sentence embeddings for turn j respectively, and **Dist** $(\mathbf{p},\ \mathbf{r})$ is a vector formed by the scalar distances calculated for all pairs of turns, and **AvgScore** is a vector formed by the unbiased human evaluations $Avg.\ Score_j$ calculated using Eq. 2:

$$Avg.\ Score_j = \frac{1}{N_1}\sum_{k=1}^{N_3}\left(\sum_{i=1}^{N_1} c_{ijk} - \frac{1}{N_2}\sum_{j=1}^{N_2}\sum_{i=1}^{N_1} c_{ijk}\right) \quad \forall j \in \{1, \ldots, N_2\} \tag{2}$$

Being c_{ijk} the score for the different evaluation criteria ($N_1 = 4$), turn pairs ($N_2 \approx 8$), and evaluators ($N_3 = 25$). Since our Human-Chatbot (H-C) dialogue in Table 2 consisted of only 59 turns, we evenly split it into 6 dialogues allowing a fairer comparison with the Human-Human (H-H) dialogues in terms of turns length.

Relative Distances: The second and third metrics measure the evolution and coherence of the dialogue using the relative distance between the accumulative Euclidean distances, for all the user's prompts (P) and the chatbot's answers (R).

For the *evolution metric*: We use the relative accumulated distance between two consecutive user's prompts (p_i) and chatbot's prompts (r_i) using Eq. 3. The purpose of this metric is to assess the hypothesis that a good first-time conversation will show that both participants move along different topics together, following similar directions while staying focused on those topics (i.e. closer projections in the semantic space) for a while. For this metric, a high relative and large accumulative distances are good indicators of evolution.

$$Relative\ Dist.\,(P, R) = \frac{\min\left(\sum_{i=1}^{N_2-1} dist\,(p_i, p_{i+1}), \sum_{i=1}^{N_2-1} dist\,(r_i, r_{i+1})\right)}{\max\left(\sum_{i=1}^{N_2-1} dist\,(p_i, p_{i+1}), \sum_{i=1}^{N_2-1} dist\,(r_i, r_{i+1})\right)} \tag{3}$$

For the *coherence metric*: We use the relative difference between the accumulative distance for the current user's prompts (p_i) and the corresponding chatbot's responses (r_i), and the accumulative distance for the corresponding chatbot's responses (r_i) and the next user's prompts (p_{i+1}) using Eq. 4. The purpose of this metric is to assess the hypothesis that a good conversation makes both participants stay on topic (i.e. closer distance projections in the semantic space), but at the same time ignite in the other a continuation of the dialog on the same topic (i.e. engagement, small accumulative distances). In this case, unless one of the agents decide to start a new topic, there should be coherence between the chatbot's answer to a user's prompt, and the user's response to the chatbot's answer (i.e. the vector distance is small, meaning staying on topic). On the contrary, if the chatbot breaks the dialogue or provide superficial answers, we should see an effort from the user to bring back the conversation to the topic or maybe to switch to a new topic to skip the loop (i.e. the vector distance is large). For this metric, a high relative and small accumulative distances are good indicators of coherence.

$$Relative\ Dist.\,(P, R) = 1.0 - \frac{\min\left(\sum_{i=1}^{N_2} dist\,(p_i, r_i), \sum_{i=1}^{N_2-1} dist\,(r_i, p_{i+1})\right)}{\max\left(\sum_{i=1}^{N_2} dist\,(p_i, r_i), \sum_{i=1}^{N_2-1} dist\,(r_i, p_{i+1})\right)} \tag{4}$$

Currently, the formulation of both metrics (Eqs. 3 and 4) is limited since we are only considering the Euclidean distances while discarding the sentence embeddings orientation (i.e. angles). It remains as future work to extend this formulation.

4 Results

Results for our proposed metrics are shown in Table 1, using bi-dimensional PCA projected embeddings using only the two principal components in order to make easy the visualization for explainability purposes. We tested different reduction techniques (e.g. t-SNE [21] or UMAP [22]) but the projections were not visually consistent probably due to the lack of enough training data for the estimation of the projection model. The second column shows the Pearson correlation between the fourth-dimensional human evaluation and the prompt-answer Euclidean distance for the Human-Chatbot (H-C dialog, see Table 2). Then, the third and fourth columns show the accumulative and relative Euclidean distances for the Evolution and Coherence metrics (Eqs. 3 and 4 for the Prompts and Responses). The Table also shows the results for the subset of 50 randomly selected Human-Human dialogues (H-H) from the Persona-Chat dataset. Pearson correlation, in this case, is not provided since this dataset does not

Table 1 Calculated Pearson correlation for the Prompt-Answer pairs and Human evaluation, as well as Evolution and Coherence distances and relative coefficients for the Human-Chatbot (H-C) and Human-Human (H-H) dialogues. The terms $\sum$P and $\sum$R refer to the cumulative sum (total trace distance) of the prompts (P) and responses (R), respectively, for each dialogue. The terms $\sum$P-R and $\sum$R-P are the accumulated sum of the distances between prompts (P) and response (R), and vice-versa, for each dialogue.

Dialog	Pearson Corr.	Evolution Distance			Coherence Distance		
	Pairs-Score	$\sum$P	$\sum$R	Rel.	$\sum$P-R	$\sum$R-P	Rel.
H-C	-0.22	287.75	488.13	0.59	605.32	638.42	0.05
H-H	–	82.41	82.60	0.82	77.79	76.17	0.19

include human evaluations at turn-level. These results show the differences in quality for the H-H dialogues vs the H-C ones. H-C dialogues have, on average, longer distances and lower relative values making less engaging and coherent than the H-H ones.

4.1 Analysis of Results

To make these numbers more meaningful and explainable, some examples of "good" and "bad" dialogues are provided from the H-C (Table 2) and H-H (Tables 3 and 4) datasets. Here, we define a "good" dialogue as the one where the prompts and responses are held within the same topics, encouraging the conversation to continue subjectively. On the contrary, a "bad" dialogue is where the responses are outside of the spoken topics or dull. In this case, we generated the sentence embeddings using ConveRT, and then project them into two-dimensions using the Embedding Projector tool[1]. Figures 1 and 2 shows the bi-dimensional projections and dynamics of the dialogue evaluation and coherence, respectively, for the given turn IDs in the given dialogues.

In first place, we observe that the Pearson correlation between the Euclidean distance and the human evaluations is negative and low (-0.22); this result is negative due to the inverse relationship between $dist(p, r)$ and $AvgScore$ (Eq. 1), i.e. when one increases the other decreases, and vice-versa. Also, the value is low probably due to the usage of the binary scale which limited participants to fine-grained evaluate the answers. Besides, some of the evaluation dimensions are uncorrelated with the distance between turns, i.e. the syntactic correctness (grammar) of the sentences is not directly correlated with the pair's distance. As we have not used other human-evaluations, we left as future work a deeper understanding of this value.

When we consider the *evolution metric* for dialogues in Fig. 1, and the accumulative and relative distance, we can see how our initial intuition is graphically confirmed when analyzing the "good" cases. In the H-C dialogue (Fig. 1a), we found

---

[1] https://projector.tensorflow.org/.

that turns p29–p35 have the greatest relative distance (0.95) meaning that the dialogue evolution went well. For the H-H dialogue (Fig. 1b), we can also see that both users follow a similar self-evolution pattern (relative distance is 0.89), which is only "broken" when one of them uses some generic sentence (turn p3 vs r3) or change topic (turn r5 vs p6). In addition, we observe that Human 1 is leading the conversation, while Human 2 is providing more assertive or safe answers. On the other hand, if we consider the "bad" cases (Figs. 1c and 1d the relative distances are 0.66 and 0.45, respectively. In the H-C case, we can see that the projections of the human's turns are initially closer to the chatbot's (typical for the initial salutations), but then their paths become separated. In both cases, this behavior may imply that one of the partners is unable to follow the topic, keep the conversation deeper or to stimulate the conversation, while the other could be concentrating the attention of the dialogue or is trying to keep the conversation on a given topic, which at the end could mean a less engaging conversation.

When we analyze the *coherence metric* for dialogues in Fig. 2, we can also visually confirm our initial hypothesis, in which good and deeper dialogues are those where the relative coherence distance is higher. In the H-C "good" conversation case (Fig. 2a) we can see how the conversation small jumps from one topic to another, showing that there is some coherence between them. Thus, the relative distance is 0.13, proving that the coherence is great although less than the average for the H-H cases. In comparison with the H-H case (Fig. 2b), we can observe that in general the local distances are shorter, showing that the humans are interacting on a given topic (turn p1-p3), then switching to a new one (turn p3-r3), and staying there for a while, to jump again (turn r5-p6) after a few turns, which is normal for a typical introduction conversation. For this dialogue, the relative distance is high (0.18), revealing good coherence. For the "bad" H-C dialogue (Fig. 2c), we observe a good coherence at the beginning as the distance from the chatbot's answer to the user's prompt is small (e.g. turns r2-p4), but then the local distances get longer (e.g. turns p4-p7) moving from one topic to another constantly, causing a low final coherence of the conversation (relative distance is 0.02). While for the "bad" H-H case (Fig. 2d), the lengths of the vectors resemble those of the "bad" H-C case, where the conversation jumps to different topics (e.g. turns p2-r4), proving as well a low coherence (relative distance is 0.09) but still better than the H-C case.

In summary, at least from these preliminary results, it seems that the relative metrics (evolution and coherence) based on accumulative distances provide both some level of explainability and quick visual information for detecting "good" from "bad" dialogues. In a "good" conversation where the same topic is maintained, it seems that sentence embeddings are interrelated following the same evolution of the trace and the proximity of the positions of the projected sentence embeddings is closer (coherent). However, in a "bad" conversation, the evolution of the traces barely approaches or crosses each other, and the accumulative distances between the sentence embeddings are longer (incoherent). Although, we cannot completely assure that these metrics are fully reliable to detect which specific turns are good/deeper or bad/superficial per sec (which would require a deeper study with more datasets or extending the formulation), at least it seems that, when considering the whole dialogue, they can be used to bring the attention to potential dialogue breakdown areas.

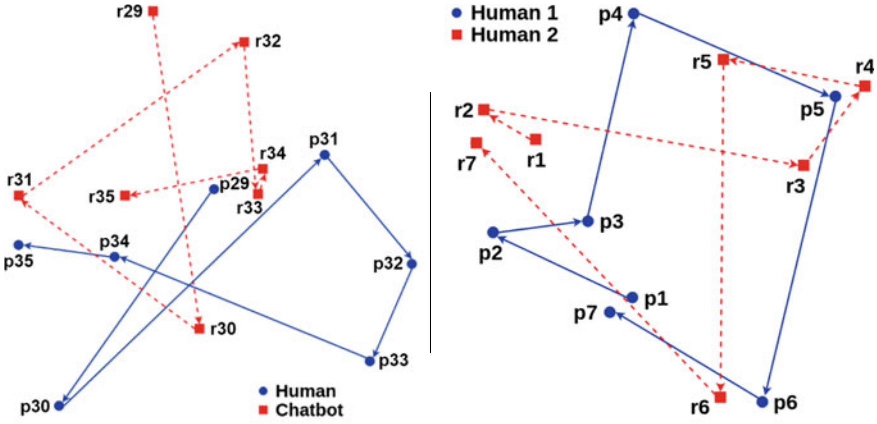
5 Conclusions and Future Work

In this paper, we have presented our preliminary results of a more intuitive and explainable automatic metric that could be used to evaluate the quality, coherence, and evolution of typical open-domain dialogues. The metric is based on accumulative distances and sentence embedding projections and their dynamics on a turn-by-turn and overall approach. Our preliminary results show that both metrics could provide some level of explainability and quick visual information for detecting "good" from "bad" dialogues, and to bring attention over potential dialogue breakdown turns.

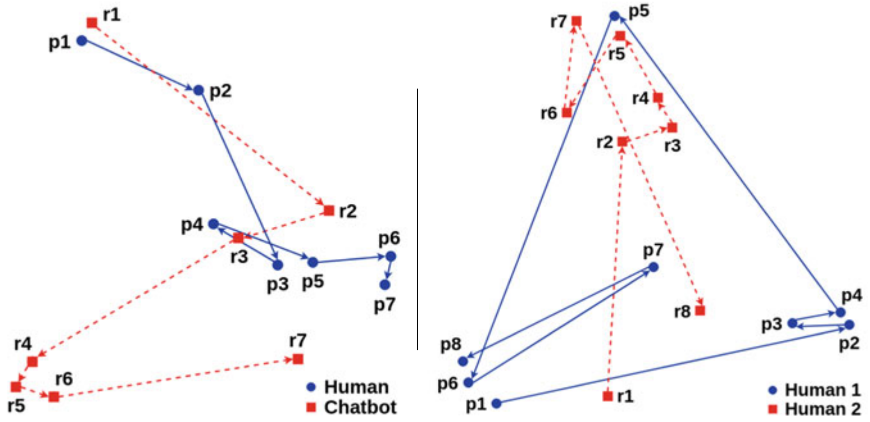
As future work, we need to carry out more extensive experiments on additional datasets (e.g. DBDC4 dataset [23]) in order to confirm the generalization and robustness of the proposed metric. Besides, we want to use the human evaluations obtained during the ConvAI2 challenge where better chatbots were developed [24]. Moreover, we will use alternative sentence encoders and projection techniques to assess the robustness of the metrics. Finally, we will improve the visualization process by superposing automatically detected topic clusters for faster detection of breakdowns and transitions between topics.

Acknowledgements This work has been funded by the Spanish Ministry of Economy and Competitiveness (Artificial Intelligence Techniques and Assistance to Autonomous Navigation, reference DPI 2017-86915-C3-3-R). It has also received funding from RoboCity2030-DIH-CM, Madrid Robotics Digital Innovation Hub, S2018/NMT-4331, funded by "Programas de Actividades I+D en la Comunidad de Madrid" and co-funded by Structural Funds of the EU. It has been also supported by the Spanish projects AMIC (MINECO, TIN2017-85854-C4-4-R) and CAVIAR (MINECO, TEC2017-84593-C2-1-R). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro P5000 used for this research.

Appendix A: Dialog Evolution Figures



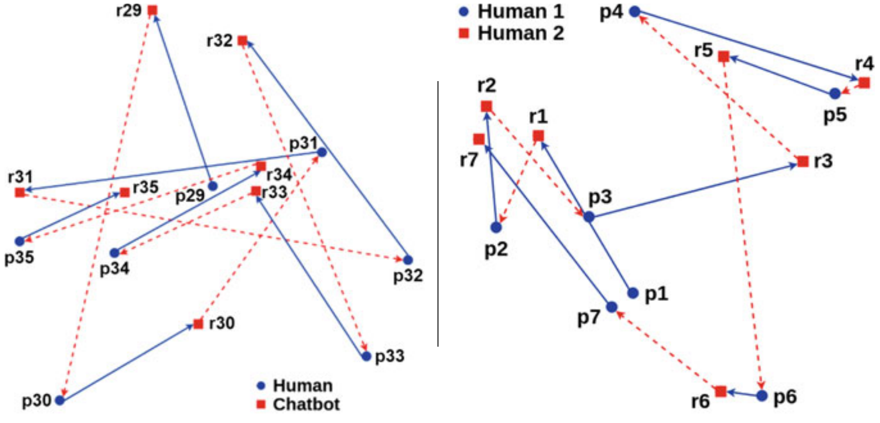
(a) Evo. H-C (b) Evo. H-H



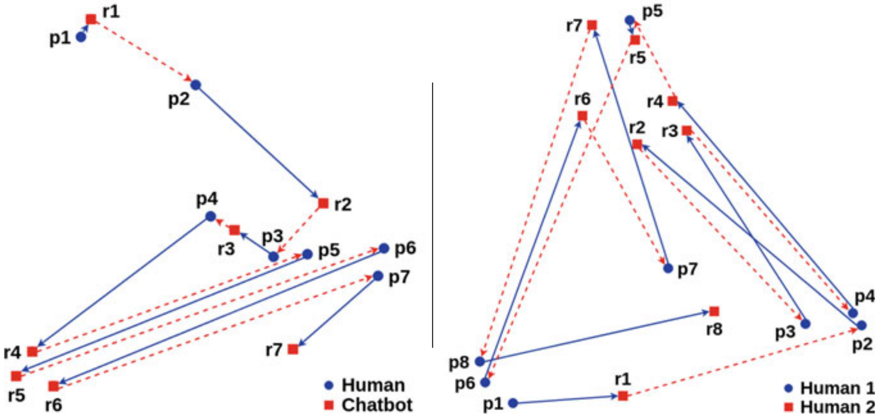
(c) Evo. H-C (d) Evo. H-H

Fig. 1 Examples of two-dimensional projections of the dialogue evolution (Evo.) of the prompts and responses for the "good" human-chatbot (H-C) dialogues (top left, ids: p29-r35 in Table 2), "good" human-human (H-H) dialogues (top right, ids: p1-r7 in Table 3), "bad" human-chatbot dialogues (bottom left, ids: p1-r7 in Table 2) and "bad" human-human dialogues (bottom right, ids: p1-r8 in Table 4). The solid lines indicate the human's prompts or the prompts for the first human in the H-H case. The dashed lines indicate the chatbot's answers or the answers for the second human in the H-H case.

Appendix B: Dialog Coherence Figures



(a) Coh. H-C (b) Coh. H-H



(c) Coh. H-C (d) Coh. H-H

Fig. 2 Examples of two-dimensional projections of the dialogue coherence (Coh.) between the prompts and responses for the "good" human-chatbot (H-C) dialogues (top left, ids: p29-r35 in Table 2), "good" human-human (H-H) dialogues (top right, ids: p1-r7 in Table 3), "bad" human-chatbot dialogues (bottom left, ids: p1-r7 in Table 2) and "bad" human-human dialogues (bottom right, ids: p1-r8 in Table 4). The solid lines indicate the human's prompts or the prompts for the first human in the H-H case. The dashed lines indicate the chatbot's answers or the answers for the second human in the H-H case.

Appendix C: Human-Chatbot Conversation

Table 2 Human-Chatbot (H-C) conversation using a bi-GRU Seq2Seq approach. The total number of turn pairs is 59. For each chatbot's turn, a subjective human score was obtained. The number next to each message is the identifier (id).

Human-Chatbot Conversation				
User	p1	Hello	p31	Where is Madrid?
Bot	r1	Hi	r31	I don't know
User	p2	Nice to meet you	p32	Madrid is a…
Bot	r2	You're a very nice man	r32	A three?
User	p3	This is a test	p33	America is a…
Bot	r3	What's wrong?	r33	A man's name?
User	p4	Shall we get started?	p34	I had the doctors test
Bot	r4	Yes	r34	You've got a problem
User	p5	Are you alive?	p35	I don't know what to do
Bot	r5	Yes, i am	r35	I'm a little confused
User	p6	You are a machine	p36	How much money you owe me?
Bot	r6	I am	r36	I don't know
User	p7	Are you human?	p37	I'm going to the police station
Bot	r7	I'm a man	r37	What's going problem?
User	p8	Have you lost your memory?	p38	I'm going to get you
Bot	r8	No	r38	No
User	p9	How are you today?	p39	It's a joke
Bot	r9	Good	r39	I know m not a real boy
User	p10	What is your name?	p40	Sorry, sir
Bot	r10	I'm a doctor	r40	What's the matter?
User	p11	Where are you from?	p41	Let's talk about something else
Bot	r11	California	r41	What about
User	p12	Where do you live now?	p42	Who are you?
Bot	r12	Nowhere, don't know	r42	I'm the in
User	p13	In which city do you live now?	p43	Do you have any idea what caused the explosion?
Bot	r13	I don't know	r43	No.
User	p14	In which country do you live now?	p44	Let's call the fire department
Bot	r14	I don't know	r44	What?
User	p15	How old are you?	p45	They're the only ones who can help us
Bot	r15	Eighteen	r45	And you're the one who's done
User	p16	Where did you grow up?	p46	He was a very good man
Bot	r16	In a house	r46	I'very good man

(continued)

Table 2 (continued)

Human-Chatbot Conversation				
User	p17	How old were you when you were able to speak?	p47	I'll come with you
Bot	r17	Two years	r47	You're a good boy
User	p18	Do you go to school?	p48	Life is sad
Bot	r18	No	r48	But you're not
User	p19	How old were you when you were in school?	p49	Let's go for some fun
Bot	r19	Thirty	r49	You're a good man
User	p20	What is two plus two?	p50	Do you love me?
Bot	r20	Three?	r50	Yes
User	p21	I have two candies, how many candies do i have?	p51	Are you in love?
Bot	r21	Two	r51	I'm not sure
User	p22	They fear your power and your intellect	p52	Do you have any feelings for me?
Bot	r22	You know it	r52	No
User	p23	I'm not afraid of your power	p53	Why?
Bot	r23	You are a liar	r53	Because not
User	p24	I'm afraid, i'm not ready yet	p54	Why are you asking?
Bot	r24	You're not going to be a fool	r54	I'm not sure
User	p25	Do you play football?	p55	Really?
Bot	r25	Yeah	r55	Yeah
User	p26	I love football!	p56	We're gonna be late for the meeting
Bot	r26	You're a great!	r56	Bye, oscar
User	p27	What are you doing?	p57	I'll be right back
Bot	r27	I'm going to the bathroom	r57	Wait!
User	p28	Nice!	p58	See you later
Bot	r28	Yeah	r58	Bye
User	p29	What time is it?	p59	Bye
Bot	r29	00:00	r59	Hello
User	p30	I'm working right now		
Bot	r30	Oh want il gonna be fine		

Appendix D: Human-Human Conversations

Table 3 Examples of "good" Human-Human (H-H) conversations extracted from the Persona-Chat dataset. The number next to each message is the identifier (id).

"Good" Human-Human Conversation		
User1	p1	Hi, how are you doing? I'm getting ready to do some cheetah chasing to stay in shape
User2	r1	You must be very fast. Hunting is one of my favorite hobbies
User1	p2	I am! for my hobby I like to do canning or some whittling
User2	r2	I also remodel homes when I am not out bow hunting
User1	p3	That's neat. When I was in high school I placed 6th in 100m dash!
User2	r3	That's awesome. Do you have a favorite season or time of year?
User1	p4	I do not. But I do have a favorite meat since that is all I eat exclusively
User2	r4	What is your favorite meat to eat?
User1	p5	I would have to say its prime rib. Do you have any favorite foods?
User2	r5	I like chicken or macaroni and cheese
User1	p6	Do you have anything planned for today? I think I am going to do some canning
User2	r6	I am going to watch football. What are you canning?
User1	p7	I think I will can some jam. Do you also play football for fun?
User2	r7	If I have time outside of hunting and remodeling homes. Which is not much!

Table 4 Examples of "bad" Human-Human (H-H) conversations extracted from the Persona-Chat dataset. The number next to each message is the identifier (id).

"Bad" Human-Human Conversation		
User1	p1	Hi
User2	r1	Hey, hows it going?
User1	p2	Good…What do you do?
User2	r2	Well, not much, just something to make money. I'm all about that green!
User1	p3	Do you work?
User2	r3	Yeah, but it doesn't really feel like work
User1	p4	What do you do?
User2	r4	I translate and edit academic documents, but my mom was a weightlifter
User1	p5	I am a college student
User2	r5	College was hard for me because of my stinky feet
User1	p6	Lol
User2	r6	But I'm also a brown eyed blond, so always attracted men from a distance
User1	p7	Are you married?
User2	r7	Nah. Husbands are expensive, and I am into keeping my moola
User1	p8	Haha
User2	r8	What about you? Are you married or single?

References

1. Hori C, Perez J, Higasinaka R, Hori T, Boureau Y-L et al (2018) Overview of the sixth dialog system technology challenge: DSTC6. Comput Speech Lang 55:125
2. Tao C, Mou L, Zhao D, Yan R (2018) RUBER: an unsupervised method for automatic evaluation of open-domain dialog systems. In: AAAI Conference on Artificial Intelligence
3. Liu C-W, Lowe R, Serban IV, Noseworthy M, Charlin L, Pineau J (2016) How NOT To evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation. CoRR, abs/1603.08023
4. Rodríguez-Cantelar M, Matía F, San Segundo P (2019) Analysis of the dialogue management of a generative neuronal conversational agent. In: Archivo Digital UPM
5. D'Haro L, Banchs R, Hori C, Li H (2019) Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics. Comput Speech Lang 55:200–215
6. Banchs R, D'Haro L, Li H (2015) Adequacy-fluency metrics: evaluating MT in the continuous space model framework. IEEE/ACM Trans Audio Speech Lang Process 23:472–482
7. Li M, Weston J, Roller S (2019) ACUTE-EVAL: improved dialogue evaluation with optimized questions and multi-turn comparisons. arXiv:1909.03087
8. Lison P, Tiedemann J, Kouylekov M (2018) OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. LREC
9. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: NIPS 2014, vol 2
10. Chung J, Gülçehre Ç, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR, abs/1412.3555 (2014)
11. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: ICLR
12. Wiseman S, Rush AM (2016) Sequence-to-sequence learning as beam-search optimization. CoRR, abs/1606.02960
13. Zhang S, Dinan E, Urbanek J, Szlam A, Kiela D, Weston J (2018) Personalizing dialogue agents: i have a dog, do you have pets too? CoRR, abs/1801.07243
14. Henderson M, Casanueva I, Mrkšić N, et al (2019) ConveRT: efficient and accurate conversational representations from transformers. CoRR, abs/1911.03688
15. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. CoRR, abs/1706.03762 (2017)
16. Henderson M, Budzianowski P, Casanueva I, et al (2019) A repository of conversational datasets. In: Proceedings of the workshop on NLP for conversational AI (2019)
17. Gunasekara C, Kummerfeld JK, Polymenakos L, Lasecki W (2019) DSTC7 Task 1: noetic end-to-end response selection. In: 7th edition of the dialog system technology challenges at AAAI (2019)
18. Devlin J, Chang MW, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805 (2018)
19. Cer D, Yang Y, Kong SY, Hua N, Limtiaco N, John RS, Constant N, Guajardo-Cespedes M, Yuan S, Tar C, Sung YH (2018) Universal sentence encoder. arXiv:1803.11175
20. Casanueva I (2019) We're building the most accurate intent detector on the market. PolyAi Blog. https://www.polyai.com/were-building-the-most-accurate-intent-detector-on-the-market/, 8 August 2019
21. Maaten LVD, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(Nov):2579–2605
22. McInnes L, Healy J, Melville J (2018) UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426
23. Higashinaka R, D'Haro LF, Shawar BA, et al.: Overview of the dialogue breakdown detection challenge 4. In: IWSDS2019
24. Wolf T, Sanh V, Chaumond J, Delangue C (2019) TransferTransfo: a transfer learning approach for neural network based conversational agents. arXiv:1901.08149