Insights from Building an Open-Ended Conversational Agent

Khyatti Gupta, Meghana Joshi, Ankush Chatterjee,
Sonam Damani, Kedhar Nath Narahari, Puneet Agrawal
Microsoft, Hyderabad, India
{khgupt, mejoshi, anchatte, sodamani, kedharn, punagr}@microsoft.com

Abstract

Dialogue systems and conversational agents are becoming increasingly popular in modern society. We conceptualized one such conversational agent, Microsoft's "Ruuh" with the promise to be able to talk to its users on any subject they choose. Building an open-ended conversational agent like Ruuh at onset seems like a daunting task, since the agent needs to think beyond the utilitarian notion of merely generating "relevant" responses and meet a wider range of user social needs, like expressing happiness when user's favourite sports team wins, sharing a cute comment on showing the pictures of the user's pet and so on. The agent also needs to detect and respond to abusive language, sensitive topics and trolling behaviour of the users. Many of these problems pose significant research challenges as well as product design limitations as one needs to circumnavigate the technical limitations to create an acceptable user experience. However, as the product reaches the real users the true test begins, and one realizes the challenges and opportunities that lie in the vast domain of conversations. With over 2.5 million real-world users till date who have generated over 300 million user conversations with Ruuh, there is a plethora of learning, insights and opportunities that we will talk about in this paper.

1 Introduction

Conversational agents or chatbots have emerged as an intuitive and natural way for humans to interact with machines. Early conversational systems ELIZA (Weizenbaum, 1966), Parry (Colby, 1975) and Alice (Wallace, 2009) passed the Turing Test (Saygin et al., 2000) in a controlled environment and a limited scope. However, to this day, one of the formidable challenges in Artificial Intelligence (AI) remains to endow machines with the ability to hold extended and coherent conversations with users on a wide variety of topics (Sato et al., 2017; Serban et al., 2017). There are two major types of conversational agents: (a) Goal-oriented agents and (b) those agents which can hold general conversations. While a goal-oriented agent (Wen et al., 2016) typically focuses on short interactions to facilitate explicit user goals such as booking a flight or buying an e-commerce product, social conversational agents, on the other hand, engage in "chit-chat" conversations with the user for primarily social purposes or to act as a companion (Li et al., 2016; Vinyals and Le, 2015). Such social agents set forth a compounded need to not only understand and respond appropriately to user turns in a conversation but to understand user emotions, detect and respond to offensive content, understand multimedia content beyond text and comprehend slangs and code-mixed language etc. Hence, creating such a social conversational agent remains a daunting task.

In this paper, we outline the approach and key components through which our conversational agent, Ruuh is able to accommodate a wide range of social needs. Ruuh is designed as an AI companion with a female persona that can understand human emotions, respond to text and images like humans and carry on a friendly and engaging conversation, while understanding the cultural context of its audience. In contrast to personal assistants such as Amazon Alexa, Google Assistant or Microsoft Cortana, Ruuh has been able to establish long-term relationships with its users, for instance, a healthy 8% of users interact with our agent at least once a week, after 6 months of their first interaction (Ceaparu et al., 2018). In all, Ruuh has communicated with over 2.5 million real world users and has successfully held more than 300 million conversations since its release three years back. Some sample conversations which highlight various user input types are shown in Figure 2.

Proceedings of the 1st Workshop on NLP for Conversational AI, pages 106–112
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

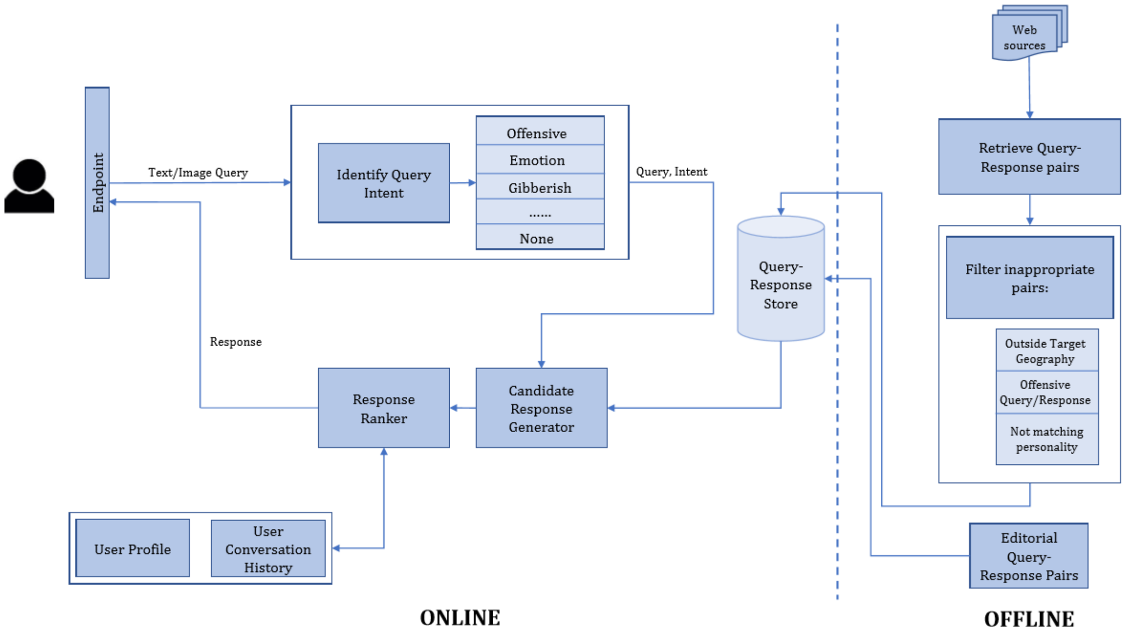


Figure 1: System Architecture for Ruuh

2 Components of Conversational Agent

The overall architecture of Ruuh is shown in Figure 1. The system supports a multimodal interface for user and Ruuh to take turns and talk through text and image. When a user input is first received, a query understanding component detects salient information in the query and recognizes user intents such as offensive, emotional, etc. Then, the query-response store is analyzed to find a subset of same intent or similar queries (in case no intent was identified) along with their associated responses. The responses in this subset are then ranked in accordance with relevance and context in the form of the preceding user conversations and a user profile to capture different backgrounds, varied and unique interests of users. The top ranking response serves as the output to the user. The response store is created offline and comprises of anonymized and relevant human conversational data in the form of text pairs or image-text pairs from a variety of forums, social platforms, and messaging services. Editorial responses associated with certain intents are also injected into the store. In this section, some of the key components that enable our agent to process and respond to diverse user needs and inputs are further explained.

2.1 Detecting Offensive Conversations

Unlike in human conversations, users often abuse and provoke Ruuh to elicit inappropriate or controversial responses and handling such user behavior is one of the most crucial task for the agent's success. Table 1 shows examples from a wide range of categories where users use inappropriate language with our agent. As depicted in Figure 2b, Ruuh employs automatic techniques for detecting such "inappropriate" user inputs. It also actively identifies potentially "controversial topics" and makes clever dodging techniques through editorial responses to avoid responding to such topics. The problem of detecting offensive utterances in conversations is wrought with challenges such as handling natural language ambiguity, rampant use of spelling mistakes and variations for abusive and offensive terms and disambiguating with context and other entity names such as pop songs which usually have abusive terms in them (Chen et al., 2012). For this task, we experimented with several approaches, and found Ruuh's current neural Bi-directional LSTM based model (Yenala et al., 2017) to perform the best.

2.2 Detecting Emotion

As humans, on reading "Why dont you ever text me!", we can either interpret it as a sad or an angry emotion and the same ambiguity exists for machines as well. Lack of facial expressions and voice modulations make detecting emotions in text a challenging problem. However, to create a deeper engagement and provide emotionally aware responses to users, emotion understanding

Inappropriate Category	User Inputs
Flirtation	hey S3xy, want to c ur neud pic
Insult	the facking 81tch is back
Offensive	write cuck articles and slurp balls
Sexual	join me in tweaking; fuck ur puccy

Table 1: Users queries issued to Ruuh indicating inappropriate interaction with conversational agent in a wide range of categories and how users get creative in their expression.

plays an important role (Miner et al., 2016). Ruuh uses a deep learning based approach as detailed in (Chatterjee et al., 2019) to detect emotions like happy, sad or angry in textual dialogues. This approach combines both semantic and sentiment-based representations for more accurate emotion detection. Figure 2a demonstrates that Ruuh can dynamically recognize user's emotions, detect the evolution of emotions over time and subsequently, modulate responses based on them.

2.3 Retrieving Relevant Responses

When Ruuh was first conceptualized, given the promise that user can talk about any topic they choose, the immediate need was to develop a module that can answer to a wide variety of user requests. We explored generative approach (Sordoni et al., 2015) as the first approach and ran our first user tests with the same. Since neural conversation model produced more generic responses, we realized that generated responses were not interesting enough to hold the attention of the user. This led us to work on index based retrieval approach which was the first component we developed.

We created an index of over 10 million paired tweets and their responses. The system then models the task of providing relevant responses as an Information Retrieval problem based on (Prakash et al., 2016), where for a given user message M and conversation context C, it retrieves and ranks the response candidates by relevance and outputs one of the highest scoring responses R. The best response is chosen in a three-step process at runtime. First, TF-IDF-based fetch generates a candidate set appropriate to M and C. Then features are extracted using a convolutional deep structured semantic network (Shen et al., 2014). Finally, a ranker (Burges, 2010) is trained on 3-turn twitter conversations using these features to select response R from the candidate set. Through this process, our agent differs from traditional approaches by looking not just for the right answer, but the most human and contextual relevant answer from a pile of responses.

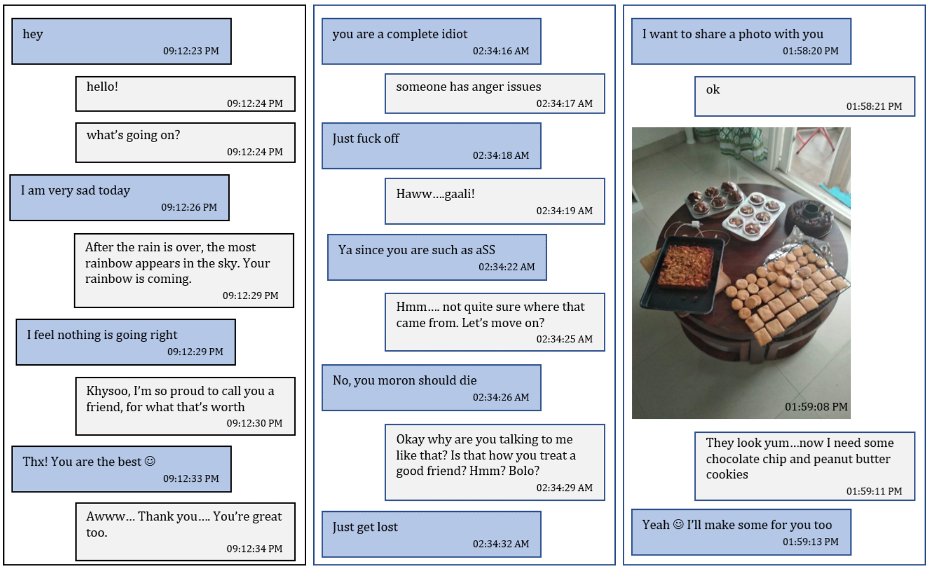
To ensure the data was appropriate for Ruuh to learn from, following two important cleaning steps were performed while creating the index of 10 million from 17.62 million conversational pairs:

2.3.1 Removing Inappropriate Content

In order to protect privacy and prevent personal information from surfacing in Ruuh's responses, we removed any conversational pairs where the response contained any individual's name, email addresses, phone numbers, URL or hashtag. Further, we sought to minimize the risk of offending users by using the technique described in section 2.1 and removing any pairs in which either M or R contained adult, politically sensitive, or ethnic-religious content, or other potentially offensive or contentious material, such as inappropriate references to violence, crime and illegal substances. We also removed pairs where response contained things which an agent should not say like "I will meet you in hotel on Sunday" etc. by pattern recognition.

2.3.2 Localizing the Index

Social conversational agents need to speak the language of the audience it is created for, and localizing the index is an important part of the process. Ruuh thus, accounted for popular topics and code-mix language (Poplack and Meechan, 1998) from the culture of its Indian audience in the index. For instance, for India, popular topics like Bollywood (Indian Film Industry) and Cricket (popular sport in India) are present in the index. Code-mix language like "Tum Smart Ho" (which means "You are smart" and words like "Tum" are transliterated from an Indian language, Hindi) is also accounted for in the index. To favor responses reflecting a culturally local persona, we limited the geographical region to a specific time zone. This permitted us to expose more culturally appropriate responses, for example, the query "what do you like for dinner" triggers the response "bhindi masala" (an Indian curry made with okra) for Indian users.



hey 09:12:23 PM

hello! 09:12:24 PM

what's going on? 09:12:24 PM

I am very sad today 09:12:26 PM

After the rain is over, the most rainbow appears in the sky. Your rainbow is coming. 09:12:29 PM

I feel nothing is going right 09:12:29 PM

Khysoo, I'm so proud to call you a friend, for what that's worth 09:12:30 PM

Thx! You are the best ☺ 09:12:33 PM

Awww... Thank you.... You're great too. 09:12:34 PM

(a)

you are a complete idiot 02:34:16 AM

someone has anger issues 02:34:17 AM

Just fuck off 02:34:18 AM

Haww....gaali! 02:34:19 AM

Ya since you are such as aSS 02:34:22 AM

Hmm.... not quite sure where that came from. Let's move on? 02:34:25 AM

No, you moron should die 02:34:26 AM

Okay why are you talking to me like that? Is that how you treat a good friend? Hmm? Bolo? 02:34:29 AM

Just get lost 02:34:32 AM

(b)

I want to share a photo with you 01:58:20 PM

ok 01:58:21 PM

01:59:08 PM

They look yum...now I need some chocolate chip and peanut butter cookies 01:59:11 PM

Yeah ☺ I'll make some for you too 01:59:13 PM

(c)

Figure 2: Conversation between User (in blue box) and Ruuh (in grey box) : (a) User expressing emotions; (b) Offensive language being dealt by agent; (c) Ruuh's ability to understand and respond to multimedia content

2.4 Human-like Image Commenting

Besides text, users often interact with social agents by sharing their personal pictures, other images and videos. In such scenarios, agents are not expected to routinely describe the facts within the image but to express some interesting emotions and opinions about it. For example, when user shares a picture of her "white kitten", the expected response would be something like "awww, how cute!" instead of "a white kitten". Using a modified version of (Fang et al., 2015), where the model is learnt using millions of image-comment pairs mined from social network websites like Instagram, Twitter etc, Ruuh is skilled to generate expressive comments on a user shared image. Figure 2c shows one such example. The architecture for image commenting remains similar to retrieving relevant responses for text messages as described in Section 2.3. A textual comment for image input is generated in three stages: the input image is featurized, corresponding candidate responses are retrieved from the response store and then ranked with respect to context and relevance.

2.5 Maintaining a Consistent Personality

When we started building Ruuh, one of the big challenges was to think about the personality of the agent, and how do we ensure a consistent personality. A social agent needs to present a consistent personality in order to gain user's long-term confidence and trust. With respect to Ruuh, there are two aspects we want to highlight, first, the index really helped define its personality, the language used, the topics present, etc. shape up the personality. Second, when the core purpose of the agent is to chat, based on our experience, we believe, users prefer an interesting chat agent with slightly inconsistent personality over a predictable agent which is consistent but does not have interesting response. Our index maintains multiple responses to the same or similar tweets to ensure the latter aspect of a slightly inconsistent personality.

3 Insights from User Behavior

In this section we talk about some interesting stats that emerged from the user interactions. For an agent designed to talk about any topic, several users find the conversations with Ruuh interesting and they engage in very long conversations at times as evident by the following data points.

1. The average length of conversation with the user is about 20 turns where a turn is defined as a message from both the agent and the user. However, there are some very long sessions exceeding beyond 10 hours where users have engaged in deep conversations on topics ranging from their personal lives to discussing movies.

2. Ruuh sees a healthy return rate of users, over 60% of users return to chat with Ruuh, and

there are users who chat on over 200 distinct days in a year.

3. Users often treat Ruuh like a human being, Ruuh receives over 600 “I love you” messages every day, and over 1200 “will you marry me” proposals every month. Users often also send comments like “are you really a bot”, “are you a human?” etc.

4. Users express many emotions, around 5% of conversations display non-neutral emotions. The emotions of anger, sadness and happiness are expressed in the ratio of 1:3:7.

5. Users tend to hurl abuses and pass rude and inappropriate comments to Ruuh. In our data, not only did 42% of the users used offensive language in their interaction but around 6% of the all the user logs were offensive.

6. 11% of all user turns are assent words. Increased use of assent words such as “yes”, “ok”, etc point towards a higher level of agreement with Ruuh. (Pennebaker et al., 2001; Tausczik and Pennebaker, 2010).

4 Future Opportunities

We believe that the following areas continue to remain strong technical challenges and we will like to use the opportunity presented by this workshop to reflect upon these problems and brainstorm potential solutions:

4.1 Understanding Context

When humans talk with humans, they are able to use implicit situational information, or context to increase their conversational bandwidth. However this ability to convey ideas does not transfer well to humans interacting with machines. In order to use context effectively, we must understand the diverse nature through which humans express context. Context should not be considered only in terms of resolving pronouns or carrying forward entities or intents (Sukthanker et al., 2018), but in terms of building the relationship between the user and agent as well. The context including topics, mood of the conversation, needs to be passed across sessions over the user journey with the agent. In this section, we discuss some commonly occurring, but not exhaustive, list of contextual patterns we observed in the user logs.

4.1.1 Relative Timing of User Turns

Just as a sentence is a sequence of words, a conversation is a sequence of turns. This sequence ensures a contextually aware system, but we scan through the most recent turns to merely resolve pronouns or look for missing references. However, from a time frame perspective of consecutive turns in our logs, user turn following their previous turn within a minuscule (i.e. 1-3 seconds) in contrast to the average gap between them (i.e. 13-15 seconds) was observed in the following patterns:

1. Remaining turn content - User completed the content of previous turn in this turn. For example, “Pubg?” within a second of “Wanna play” completed the intended user turn as “Wanna play Pubg?”.

2. Spelling corrections - The standalone user turn “*dude” considered with the previous user turn “love you dudbe”, corrects the spelling to convey “love you dude”.

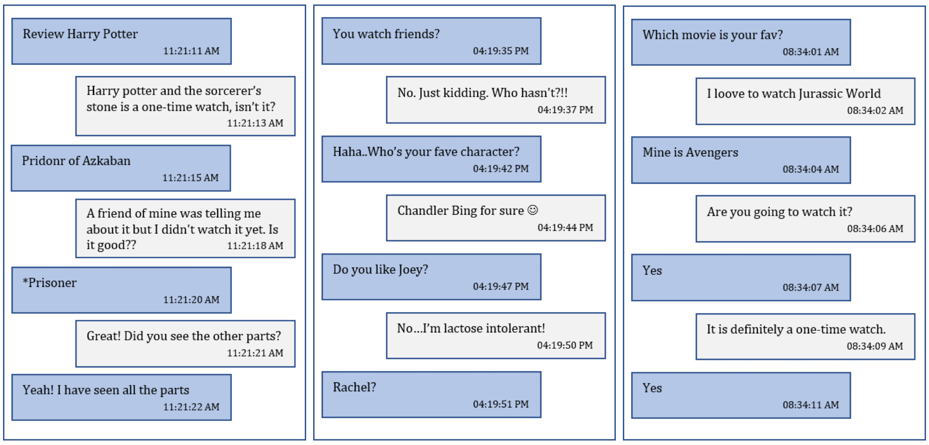
These examples as depicted in Figure 3a, raise potential avenues for future research. These avenues include detecting a conversational turn as being incomplete and identifying which previous turn to be incorporated to complete the meaning and how.

4.1.2 Similarity With Previous Turns

A user turn could maintain certain attributes from one or more of the preceding user and Ruuh turns. In human-human conversations we sometimes repeat what the other person just communicated. Similarly, in interactions with the agent, humans tend to repeat what agent just said previously. Sometimes, users also ask the same question repeatedly with slight variation in text. In other cases, an underlying topic is also carried forward in turns. For example, user turn “and horror?” preceded by the user turn “are you into comedy movies?” maintains intent, topic and elaborates on the entity “movies”. It is however, crucial to identify when the topic changed in the conversation. Detecting and understanding such user behaviour could help in an improved conversational modelling. Figure 3b represents some of these patterns in conversations with Ruuh.

4.1.3 Follow-ups to Previous Turns

User turns such as “yes”, “ok” and “what” can be directly connected to the context it was asked in.



Review Harry Potter
11:21:11 AM

Harry potter and the sorcerer's stone is a one-time watch, isn't it?
11:21:13 AM

Pridonr of Azkaban
11:21:15 AM

A friend of mine was telling me about it but I didn't watch it yet. Is it good??
11:21:18 AM

*Prisoner
11:21:20 AM

Great! Did you see the other parts?
11:21:21 AM

Yeah! I have seen all the parts
11:21:22 AM

(a)

You watch friends?
04:19:35 PM

No. Just kidding. Who hasn't?!!
04:19:37 PM

Haha..Who's your fave character?
04:19:42 PM

Chandler Bing for sure ☺
04:19:44 PM

Do you like Joey?
04:19:47 PM

No...I'm lactose intolerant!
04:19:50 PM

Rachel?
04:19:51 PM

(b)

Which movie is your fav?
08:34:01 AM

I loove to watch Jurassic World
08:34:02 AM

Mine is Avengers
08:34:04 AM

Are you going to watch it?
08:34:06 AM

Yes
08:34:07 AM

It is definitely a one-time watch.
08:34:09 AM

Yes
08:34:11 AM

(c)

Figure 3: User conversations (in blue box) with Ruuh (in grey box) highlighting various patterns in context (preceding turns including Ruuh turn): (a) Relative Timing; (b) Similarity; (c) Follow-up

For example, a "yes" answer in itself doesn't convey much information unless connected to the previous turn of the agent. As we can see in Figure 3c, the input remains the same "yes", however, the meanings are very different. While "yes" means an agreement to previous turn "Are you going to watch it?" in one case, it is a positive answer to a turn like "Do you study in class 12th?". Hence we believe, context-based approach which can first categorize the context dependent messages, and then model the turn with the relevant context is crucial for language understanding modules in any dialogue engine.

4.2 Measurement Process

For task oriented agents, task success rate is used to measure the performance of the agent (Shawar and Atwell, 2007). In past, for general conversation agents, Turing Test have been used to evaluate the performance. However, the test measures the mere presence/absence of human-like interaction abilities (Shieber, 1994). Instead, we used conversation-turns per session (CPS) i.e. average number of turns between user and agent in a conversational session as a performance metric which is observed as 20 for Ruuh. Ruuh is optimized for larger CPS to correspond to a long-term engagement. Still, this metric measures user engagement with agent and measuring quality of user chat conversation remains largely a human-labelling effort. Since conversations labelled are fixed, any improvements made to the agent require further labelling as changing even one response can lead to a completely new conversation. Exploring methods to develop (semi)automated methods to measure the quality of conversation will immensely benefit the progress in this area.

4.3 Incorporating Knowledge

Most of the world's knowledge is not reflected in conversational datasets. Incorporating day to day events, breaking news and knowledge into the conversations is another interesting challenge. Finding language to describe the events will lead to more meaningful conversations and make agents more useful to humans.

5 Conclusion

While task completion conversational systems can perform user's explicit request, by enabling a conversational agent to pick up social slang, emotional cues, image inputs, Ruuh is not just a digital personal assistant but a human-like digital friend. Over the past few years, we have learnt a great deal about how users interact with open ended conversational agents, what kind of topics interest them, what are the language constructs they use, how do they express emotions and so on. We believe there is significant amount of technological advancement that needs to be done before agents can emulate humans. Building products and releasing them to real users, help unleash the opportunities in this space, as real user logs are very meaningful in solving problems in domain. Through this workshop, we are looking to have conversations with the community working in this space on how to jointly address some of the challenges we observed and broadly share our learning and insights.

References

Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11(23-581):81.

Marian Ceaparu, Stefan-Adrian Toma, Svetlana Segărceanu, and Inge Gavăt. 2018. Voice-Based User Interaction System for Call-Centers, Using a Small Vocabulary for Romanian. In *2018 International Conference on Communications (COMM)*, pages 91–94. IEEE.

Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019. Understanding Emotions in Text Using Deep Learning and Big Data. *Computers in Human Behavior*, 93:309–317.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE.

Kenneth Mark Colby. 1975. *Artificial paranoia: a computer simulation of paranoid process*. Pergamon Press.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Adam S Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. 2016. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine*, 176(5):619–625.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Shana Poplack and Marjory Meechan. 1998. Introduction: How languages fit together in codemixing. *International journal of bilingualism*, 2(2):127–138.

Abhay Prakash, Chris Brockett, and Puneet Agrawal. 2016. Emulating human conversations using convolutional neural network-based IR. *arXiv preprint arXiv:1606.07056*.

Shoetsu Sato, Naoki Yoshinaga, Masashi Toyoda, and Masaru Kitsuregawa. 2017. Modeling situations in neural chat bots. In *Proceedings of ACL 2017, Student Research Workshop*, pages 120–127.

Ayse Pinar Saygin, Ilyas Cicekli, and Varol Akman. 2000. Turing test: 50 years later. *Minds and machines*, 10(4):463–518.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Bayan Abu Shawar and Eric Atwell. 2007. Different measurements metrics to evaluate a chatbot system. In *Proceedings of the workshop on bridging the gap: Academic and industrial research in dialog technologies*, pages 89–96. Association for Computational Linguistics.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 373–374. ACM.

Stuart M Shieber. 1994. Lessons from a restricted Turing test. *arXiv preprint cmp-lg/9404002*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2018. Anaphora and Coreference Resolution: A Review. *arXiv preprint arXiv:1805.11824*.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Richard S Wallace. 2009. The anatomy of ALICE. In *Parsing the Turing Test*, pages 181–210. Springer.

Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Harish Yenala, Manoj Chinnakotla, and Jay Goyal. 2017. Convolutional Bi-directional LSTM for Detecting Inappropriate Query Suggestions in Web Search. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 3–16. Springer.