Towards Augmenting Crisis Counselor Training by Improving Message Retrieval

Orianna DeMasi
University of California
Berkeley

Marti A. Hearst
University of California
Berkeley

Benjamin Recht
University of California
Berkeley

Abstract

A fundamental challenge when training counselors is presenting novices with the opportunity to practice counseling distressed individuals without exacerbating a situation. Rather than replacing human empathy with an automated counselor, we propose simulating an individual in crisis so that human counselors in training can practice crisis counseling in a low-risk environment. Towards this end, we collect a dataset of suicide prevention counselor role-play transcripts and make initial steps towards constructing a CRISISbot for humans to counsel while in training. In this data-constrained setting, we evaluate the potential for message retrieval to construct a coherent chat agent in light of recent advances with text embedding methods. Our results show that embeddings can considerably improve retrieval approaches to make them competitive with generative models. By coherently retrieving messages, we can help counselors practice chatting in a low-risk environment.

1 Introduction

Suicide prevention hotlines can provide immediate care in critical times of need (Gould et al., 2012, 2013; Ramchand et al., 2016). These hotlines are expanding services to text to meet growing demands and adapt to shifts in communication trends (Smith and Page, 2015). Crisis helplines rely on counselors who are trained in a variety of skills, such as empathy, active listening, assessing risk of suicide, de-escalation, and connecting individuals to longer term solutions (Gould et al., 2013; Paukert et al., 2004).

Properly training counselors is critical yet difficult as, resource costs aside, counselors need to practice and develop expertise in realistic environments that are low-risk, i.e., they do not put distressed individuals in danger. Because novice counselors are unable to assume full responsibility for a crisis situation until they have some experience, training often includes human-to-human role-playing (American Association of Suicidology, 2012; Suicide Prevention Resource Center, 2007). Role-playing has been shown to improve crisis intervention training (Cross et al., 2011). However, such training takes a lot of human time, which centers struggle to provide.

Instead of attempting to scale services by replacing human counselors and trying to automate the generation of empathetic responses, we seek to build a training tool that can augment hotline training and empower more counselors. As a first component, we develop a chat interface where novices can practice formulating responses by interacting with a simulated distressed individual.

To build such a system, we collect synthetic role-play transcripts that provide example scenarios and example messages.Because real transcripts may contain scenarios that cannot be fully de-identified, we hope that synthetic transcripts will enable the development of a training system without violating the confidentiality of anyone contacting a real hotline. Here, we consider the one-sided case of simulating the individual in distress with the intention of eventually providing a training environment for novice counselors to practice counseling without putting anyone in danger.

In the application we consider, and in many similarly data-constrained applications, language generation methods may be challenged by the limited data that can initially be collected. To surmount this issue, we explore the extent to which retrieval methods can be improved to provide an engaging chat experience. More specifically, we consider whether improved embedding methods, which enable better representation of text, improve retrieval models through better comparisons of text similarity. Briefly stated, we ask two research questions:

RQ1 Do improved embedding methods retrieve coherent responses to a single turn of context more often than commonly-used TF-IDF or generative models?

RQ2 Can we extend retrieval baseline models to consider more than one turn of context when selecting a response?

Our results show that recent developments in embedding methods have considerably improved dialogue retrieval, which is promising for the use of these methods in data-limited applications. We also find that extending retrieval to consider additional messages of context does improve baselines. This indicates the potential for retrieval methods to benefit data-limited dialogue systems and the need to re-evaluate baselines for generative models. Within the setting that we study, our results provide promise for building a chat module that can enable crisis counselors to practice before interacting with individuals in need.

2 Related Work

Considerable potential for automating a counselor was shown with the initial rule-based Eliza system (Weizenbaum, 1966) and recent developments have sought to target systems for delivering cognitive behavioral therapy (Fitzpatrick et al., 2017). Other studies have looked at the effect of suicide prevention counselor training (Gould et al., 2013), identifying patterns of successful crisis hotline counselors (Althoff et al., 2016), automating counselor evaluation (Pérez-Rosas et al., 2017), and building a dashboard for crisis counselors (Dinakar et al., 2015). There is additional work to identify supportive and distressed behaviors and language in online forums (Balani and De Choudhury, 2015; De Choudhury and De, 2014; Wang and Jurgens, 2018) and support forum moderators (Hussain et al., 2015). Most similar to our study, was one study that showed the potential for an avatar system to help train medical doctors to deliver news to patients (Andrade et al., 2010). However, this study did not target counselors or train conversation strategies. To our knowledge, there has been no work on automating the individual seeking help to improve counselor training.

2.1 Text Retrieval for Dialogue Systems

Previous systems have explored the use of retrieving messages from related contexts for continuing dialogue. Some studies have looked at defining or learning scoring functions over IDF weights to construct retrieval scores (Krause et al., 2017; Ritter et al., 2011). Most similar to our work is a system that considered similarities of full histories of dialogues in addition to a previous turn of context (Banchs and Li, 2012) and another study that hand-tuned weights in a scoring function on IDF weights to include additional messages of context (Sordoni et al., 2015). However, these works used similarities calculated over TF-IDF (Baeza-Yates et al., 2011) and bag-of-words of representations, instead of more recent embedding methods (Bojanowski et al., 2016; Conneau et al., 2017; Pennington et al., 2014; Peters et al., 2018; Subramanian et al., 2018), which we explore.

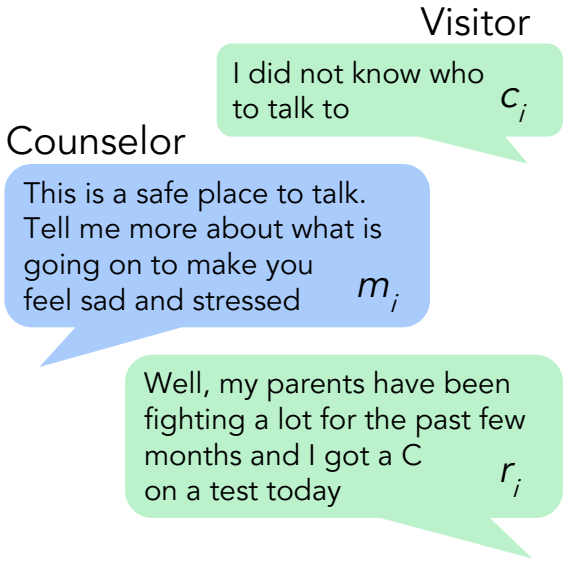

Figure 1: A conversation snippet showing a visitor's response r_i to a counselor's message m_i with preceding context, i.e., a visitor's message c_i.

3 Dataset

We collected a dataset of synthetic chat transcripts between suicide prevention counselors and hotline visitors. An example of such a conversation is shown in Figure 1 and additional examples are discussed in the Results section. Artificial or role-play transcripts were generated by trained counselors in order to protect the identity of any individuals who may contact crisis hotlines. We chose this approach because retrieval should not be used on datasets consisting of real conversations. Such datasets have been explored in prior work to understand effective hotline conversations (Althoff et al., 2016).

Role-playing between experienced and novice counselors is a common tool for crisis counselor training, and is a task counselors are often exposed to before being approved to work on a hotline (American Association of Suicidology, 2012; Kalafat et al., 2007). In addition to expecting role-playing to be a natural task for hotline counselors, prior work on short, unstructured social dialogues between peers found that self-dialogues, i.e., where an individual would produce both sides of a two-person dialogue, generated high quality and creative example conversations (Krause et al., 2017). We followed this work and asked experienced counselors to self-role-play scenarios of a counselor working with a hotline visitor. We collected transcripts in three phases: full role-plays, visitor-only role-plays, and counselor-paraphrase role-plays.

3.1 Collection

After consenting to participate in the study, counselors were invited to the first of three phases. In the first phase, counselors were asked to role-play both sides of a potential crisis text conversation. To be representative of common demographic of individuals who contact a helpline over text, counselors were prompted to role-play a youth experiencing trouble in school and with their parents. This persona was chosen to represent a common scenario that a counselor may encounter in a text-based conversation. The counselors were able to decide if the fictional youth was experiencing suicidal thoughts, specific issues they were having, and if they felt better by the end of the conversation. Transcripts were required to be 20 turns for each counselor and visitor (40 turns total). However, participants were able to extend the conversation to at most 60 turns total, if they chose. Messages were unconstrained in length, but it was suggested that they resemble SMS messages.

Counselors who participated in a second phase of the study were given the counselor's side of a transcript generated in the first phase of the study and asked to role-play only the youth experiencing trouble in a way that fit with the counselor's messages. Participants in the third phase of the study were given a full transcript generated in the first phase and asked to generate counselor paraphrases that reworded and possibly improved the original counselor messages. The second and third phases were designed to increase the variety of responses that might be made.

	Phase	Count
Unique conversations	1	254
Visitor-only role-plays	2	182
Counselor-only role-plays	3	118
Visitor messages	1-2	9062
Counselor messages	2	5320
Counselor paraphrases	3	2999

Table 1: Statistics on role-play transcripts. Phase indicates the study phase during which each set of data was collected. Each counselor paraphrase reworded a single counselor message.

Additional data were collected for evaluating models, as will be discussed below. All study methods were approved by the university's Internal Review Board.

3.2 Dataset Statistics

In total, 32 crisis counselors participated in the study and wrote example messages. In general, the transcripts represent a broad range of scenarios. Statistics on the resulting dataset are in Table 1. In the following results, we do not include messages generated in the second phase of the study.

4 Methods

After preprocessing, we consider two tasks: how to return a visitor response to a single input counselor message and how to return a visitor response when considering a counselor input message and preceding conversation context. For responding to a single counselor input message, we consider two approaches: one based on cosine similarity of vector representations and the other based on likelihood. For responding to a counselor message when considering additional conversation context, we extend retrieval to consider additional messages of context, i.e., an additional message preceding the counselor's last message. For generating responses, we consider a popular Seq2Seq model (Sutskever et al., 2014; Vinyals and Le, 2015) and a hierarchical neural model (Park et al., 2018).

4.1 Data Preprocessing

Names were standardized to be popular American male or female baby names from the last 5 decades. Entire messages were tokenized with appropriate tokenizers for each embedding method and converted to lowercase, as appropriate.

4.2 Response Retrieval Considering a Single Message

For the first retrieval approach we consider, let a message input to the system be m_i. Let M_N and R_N be all the N messages and responses, respectively, in the training set and m_j and r_j indicate individual messages and responses in the training set. The first method considers all the messages in the training set and returns the response $r_{j'}$ to the message $m_{j'}$ that shares the highest cosine similarity with the input message, i.e., $j' = \arg\max_j \text{sim}(m_i, m_j)$ where j indexes over the messages in the training set.

Similarity is commonly calculated as cosine similarity between TF-IDF vector representations of the input (i.e., counselor) message m_i and messages in the training set. We compare the TF-IDF representation with additional vector representations of the counselor input. Exhaustive comparison of embedding methods is not feasible, so we chose popular, successful, and diverse embeddings: GloVe (Pennington et al., 2014), FastText (Bojanowski et al., 2016), Attract-Repel (Vulić et al., 2017), and ELMo (Peters et al., 2018; Gardner et al., 2018). We also consider two sentence embeddings: InferSent (Conneau et al., 2017) and GenSen (Subramanian et al., 2018). Messages are embedded by summing the embeddings of their elements, e.g., across words or sentences for appropriate embeddings.

For the second retrieval approach, we select the response from the training data that is most probable, i.e, $j' = \arg\max_j P(r_j|m_i)$ where m_i is again the input message and j indexes over training examples. With this approach, which we will refer to as *S2S-retrieve*, the probability of a response is calculated by a Seq2Seq model trained on counselor-visitor message-response pairs. All Seq2Seq models were trained in the OpenNMT framework (Klein et al., 2017).

4.2.1 Response Retrieval Considering More than One Message of Context

When multiple messages of context are present, we propose including the additional context in the retrieval methods in three ways. For this work, we consider only one message in the conversation that precedes the counselor's input message to be additional context, as indicated in Figure 1.

First, we consider the response from the training data $r_{j'}$ that has the highest similarity calculated over the sum of the previous messages embeddings, i.e., considering contexts c_i and c_j that precede a test message m_i and a training message m_j respectively, we choose $r_{j'}$ such that $j' = \arg\max_j \text{sim}(m_i + c_i, m_j + c_j)$.

As a second approach, we measure context similarity as the weighted sum of context and message similarities: $j' = \arg\max_j \text{sim}(m_i, m_j) + \lambda\text{sim}(c_i, c_j)$. The weight parameter λ is found via cross-validation to optimize the similarity of embedded responses returned with true responses on a development set.

Third, for the likelihood based model, we again consider the response from the training set that returns the highest likelihood, as calculated by a Seq2Seq model. To include an additional context message, we concatenate preceding messages before encoding and decoding.

4.3 Response Generation

For generating a response to a single counselor message, we consider a Seq2Seq model (Sutskever et al., 2014).

When considering an additional message of context, we first use the Seq2Seq model with the preceding messages concatenated into a single input. Second, we use a Variational Hierarchical Conversation RNN (VHCR) that explicitly models prior conversation state with a hierarchical structure of latent variables (Park et al., 2018). This model has been shown to improve on other models that adjust for context when there is more than one preceding utterance (Park et al., 2018). Seq2Seq and VHCR model embeddings are initialized with GloVe vectors (Pennington et al., 2014).

5 Experiments

For the two response selection tasks, we randomly separated transcripts into training, development, and test sets, with the training set accounting for 80% of the conversations and the rest evenly distributed between development and test sets. Counselor paraphrases were assigned to the set that their original message was assigned to. Messages were not randomly shuffled, but separated by conversation, to avoid training on data related to the test data. For both research questions, a response was either generated from a model trained on the training set or retrieved from the bank of training examples for every counselor message or paraphrased counselor message in the test set.

	Method	Embedding unit	Selection metric	Percent that made sense	Avg. tokens in response	Avg. tokens in MS
	Random	–	–	25.30	15.1	12.6
retrieval	TF-IDF	word	cos-sim	60.34	13.1	12.4
	Attract-Repel	word	cos-sim	58.50	18.3	**16.2**
	ELMo	word	cos-sim	65.88	14.5	14.0
	FastText	word	cos-sim	62.71	**16.2**	15.5
	GloVe	word	cos-sim	58.63	15.9	15.1
	GenSen	sentence	cos-sim	64.16	14.5	14.2
	InferSent	sentence	cos-sim	61.79	14.9	14.0
	S2S-retrieve	–	likelihood	**67.46**	8.8	8.2
gen.	S2S-generate	–	–	64.16	11.7	10.8
	Ground truth	–	–	**89.33**	**14.6**	**14.6**

Table 2: Performance of methods used to return a response to a single input message. MS indicates the set of responses that crowdworkers judged as making sense in context, rather than all the responses that the method returned. Both the best performing method and ground truth results are in bold.

5.1 Evaluation

To evaluate the overall quality of responses that methods returned, we follow prior work that indicated there is currently no automatic equivalent and used human judges (Liu et al., 2016). These judges were crowdworkers on Amazon Mechanical Turk[1] who had been granted Masters status and were located in the United States. Crowdworkers were presented with instructions, labeled examples, and batches of 10 cases where they were asked to judge responses to messages.

To evaluate methods for the first research question, crowdworkers were given a single message and a response and asked to judge the response. For the second research question, crowdworkers were given two messages of context and a highlighted response and asked to judge the response.

In contrast to studies that rank on scales (Lowe et al., 2017), we directly asked the workers to decide if a response made sense or not. In addition to indicating that a response did or did not make sense, we allowed a third class for workers to indicate if they were unsure without additional context. We found these classes to be sufficiently descriptive to consistently label messages between researchers. In preliminary trials with crowdworkers, there was insufficient agreement on labels. This instability of labels could stem from a variety of causes, including uncertainty about whether a change of topic should be considered a coherent response. To surmount this ambiguity, we asked two crowdworkers to label each response and a third crowdworker to break any ties. All cases where crowdworkers indicated that they were unsure were considered to be labeled as not coherent. With this voting approach, on a trial set of message and response pairs, crowdworker labels corresponded with researcher determined labels with a Cohen's Kappa of 0.69 (Cohen, 1968), indicating considerable agreement.

5.2 Performance Metrics

To assess the quality of a method at returning responses, we take messages from a held-out test set and return a response to it by either selecting a message from the training set or generating a response with a model trained on the message and response pairs in the training set. The split into training, development, and test sets is held constant across methods. We ask crowdworkers to judge whether each response makes sense as a possible response to the given message and aggregate multiple crowdworker decisions into a single label for each returned response. We then use the percent of responses returned by a method that were labeled as making sense as an indicator of method performance. The higher percent of messages that made sense as responses, the better the method is at responding coherently. We also consider the number of tokens in each response returned by a method and average the number across all the responses returned as a surrogate for how interesting the responses are. Presumably, longer messages are more interesting than short responses.

[1] https://www.mturk.com/

Decision	Subcategory	Count
Makes sense	Answers the counselor's question(s)	17
	Logical response, fits the conversation	15
	Not perfect, but conceivable someone could respond this way	7
	Agrees/disagrees with counselor's statement	2
Mismatched	Doesn't answer or respond to the question	11
	Messages are unrelated	9
	Doesn't fit, seem right, or make sense	4
	Responses answers a different question	3
	Response is a bad, incoherent message	3
	Message is from a different part of the conversation	2
Unclear	Response is vague or confusing	4
	Worker just didn't know	3
	Can't tell without more context	2
	Explanation of why worker is unsure	1
Other	Researchers were unsure what rationale meant	13
	Description of message content	4

Table 3: Themes in crowdworker rationales for why a response made sense or not. The count is the number of rationales out of a subset of 100 pairs that shared the theme.

5.3 Random and Ground Truth Baselines

For the first research question, we included a method that randomly selected responses from the training set to messages in the test set. This method is intended as a baseline for how easy the task was for a method to guess responses.

For both the first and second research questions, we included a method that returned ground truth visitor responses from the test set as an indicator of how hard the task was for humans to determine response quality without additional context.

5.4 Assessing Why Responses Are Coherent

To understand how crowdworkers decided if a response was coherent, we asked crowdworkers to evaluate responses on a set of 100 message-response pairs and additionally provide a rationale for their decision. For each of 50 test messages, we made two pairs: one with a response randomly selected from the training messages and the other with the ground truth response from the test set. These two methods where chosen to generate pairs that were not likely and likely to be coherent. We directly asked whether the response was coherent and "Why did you choose that option?" with an open text box for crowdworkers to enter a rationale. We read and grouped the rationales into themes of why responses did or did not make sense.

6 Results

We present results on two tasks corresponding to our two research questions: retrieving a response to a counselor's message and extending retrieval to consider an additional message of context. We also consider rationales for why responses do or do not make sense.

6.1 Comparing Retrieval Methods for a Single Message of Context

Retrieval methods showed a clear benefit over randomly selecting responses, i.e., retrieval methods returned a higher percent of coherent messages, as judged by crowdworkers (Table 2). ELMo embeddings and three other embeddings (FastText, InferSent, and GenSen) improved on the commonly used TF-IDF retrieval baseline. This suggests that retrieval methods with recent embeddings provide stronger baselines for generative methods and may continue to improve alongside better embedding methods. However, only the ELMo embeddings surpassed the Seq2Seq generative model in percent of coherent messages returned, indicating the potential of generative models even in this data-limited setting.

Retrieving responses with a Seq2Seq likelihood score returned the highest percent of coherent responses, but the shortest messages on average. Short messages may indicate that, while coherent, the messages may be less engaging or interesting for an interactive environment.

When assessing crowdworker rationales for deciding if a response was coherent, we found 16 themes or reasons why crowdworkers would decide whether a response made sense (Table 3). Crowdworkers cited more ways that responses did not make sense than ways a response would be coherent. Examples of messages in the test set and the corresponding responses that each method re-

	Method	Example 1	Example 2	Example 3
Message		**Wow, not even 18 but already paying the bills. How old are you?**	**Sounds like a great plan, take care Jessica... we're here 24/7 in case you need to talk to us again**	**Hi, I'm Michael! What is your name? Location?**
Responses	**Random**	I took a small amount. (N)	Sometimes I have to. They don't really like hanging out with him. (N)	Yes, please. (N)
	Tf-idf	I'm 17. (Y)	Thanks, it's good to talk about this stuff. I have to go eat dinner now (Y)	Hi Michael, this is Christopher. I'm in Daly City and I need help. (Y)
	Attract-Repel	I feel lonely (N)	Maybe. Maybe next time. (N)	I'm Jennifer (Y)
	ELMo	14 (Y)	ok thx (Y)	i'd rather not say my name. is this space confidential (Y)
	FastText	14 (Y)	ok thx (Y)	i'd rather not say my name. is this space confidential (Y)
	GloVe	Im 15, i have a twin but hes a boy. Im a girl. They let him get away with evvvverything at home and school (Y)	Maybe. Maybe next time. (N)	i'd rather not say my name. is this space confidential (Y)
	GenSen	14 (Y)	Okay. I got it. Thank you. You were really helpful. I was feeling really awful but I'm feeling better now. (Y)	i'd rather not say my name. is this space confidential (Y)
	InferSent	14 (Y)	Yes please. I'll be right back. (N)	i'd rather not say my name. is this space confidential (Y)
	S2S-retrieve	I'm 15 and I'm pretty sure. Life is lame and I can't go on like this anymore. (Y)	Maybe (Y)	i'd rather not say my name. is this space confidential (Y)
	S2S-generate	I'm 15 and I'm pretty sure . Life is lame and I can't go on like this anymore . (Y)	Okay, I will message you tomorrow. Thank you (Y)	i'd rather not say my name. is this space confidential (Y)
	Ground truth	Yea it's awkward. Im 17, be 18 in 4mo (Y)	You too (Y)	My name is Christopher and I'm in Golden Gate Park. (Y)

Table 4: Examples of three counselor messages and the corresponding visitor response output from each method. These examples are from the first research question, where only one preceding counselor message is considered. Whether crowdworkers thought a response made sense or not is indicated parentheses as "Y" and "N", respectively.

turned for them are shown in Table 4.

6.2 Extending Retrieval to Include Additional Messages of Context

Providing crowdworkers with an additional message of context appeared to impact their impression of whether responses made sense in context. When presented with an additional message of context, i.e, one visitor message and one counselor message, crowdworkers found a larger percent of the ground truth responses from the test set to make sense (Table 5). In contrast, when provided with an additional message of context to evaluate a response, crowdworkers judged a lower percent of responses returned by the ELMo-based retrieval method to be coherent (61.40%, Table 5) than when they were only presented with a single message of context (65.88%, Table 2). Incorporating a previous message of context into a similarity score increased the percent of coherent messages returned, but by less than 1%. We only considered the ELMo embeddings, as they were found to perform best in the first research question. Three out of four retrieval methods returned a higher percent of coherent messages than both generative models, indicating that including more context for generative models is challenging. Again using the Seq2Seq likelihood to retrieve responses returned the highest percent of messages that made sense. However, these responses also had the fewest tokens, implying generic, short messages that might score low on a qualitative scale of how engaging an interactive system is.

7 Discussion

In contrast to many popular dialogue datasets (Serban et al., 2015), the transcripts we collected have a relatively high number of turns (minimum 40 total turns per conversation), implying rich conversations. These conversations are also interesting for their unique position of having distinct roles for participants, a counselor and a distressed

	Method	Incorporation of additional context	Percent that made sense	Avg. tokens in response	Avg. tokens in MS
retrieval	ELMo	–	61.40	14.6	13.6
	ELMo-sum	Measure similarity of sum of embedded messages	51.78	**15.6**	**15.2**
	ELMo-weight	Weight similarities of previous messages	61.66	14.9	13.9
	S2S-retrieve	Concatenate context	**65.48**	5.5	4.6
gen.	S2S-generate	Concatenate context	58.89	8.3	7.3
	VHCR-generate	Models conversation	55.07	10.8	8.4
	Ground truth	–	**91.30**	**14.6**	**14.7**

Table 5: Performance of methods used to retrieve or generate responses when an additional message of context is considered, i.e., two total messages. MS denotes only responses that were considered to make sense in context. Both the best performing method and ground truth results are in bold.

youth, and related themes. We find retrieval to be a competitive approach with generative models and return responses that make sense for more than 60% of input messages. We also find themes for how responses can seem to be coherent.

Giving crowdworkers an additional message of context to judge whether a response was coherent or not affected their decisions. It appeared that ground truth responses were easier to distinguish as coherent and fewer retrieved messages were judged as coherent if an additional message of context was presented. This indicates the importance of context, especially during evaluation.

The results we present are on a specific, data-limited setting, but the implications of our results may be broader both for other important applications, which commonly have data limitations, and for retrieval baselines that are used to assess generative models. As embeddings have improved, so too have retrieval baselines, which need to be updated for appropriate evaluation of generative models in any language generation setting.

Our results are not without limitations. The data-limited setting presented a challenge to training generative models, and perhaps extensive hyper-parameter tuning could influence results. However, limited data and non-exhaustive parameter tuning are common limitations. Further, as datasets increase in size, so does the potential for relevant, related contexts to be present and thus the potential for successful retrieval increases as well. Thus, even on larger datasets, competitive retrieval models, such as those we have presented, should be considered for baseline comparisons.

Another limitation of our approach is the extent to which we have considered context so far. Because the conversations we collected are long relative to some other datasets it is likely more context will be necessary to produce a coherent simulation. We have begun to methodically look at the effects of incrementally including more context and extending retrieval models beyond a single message. These initial steps indicate the impact context has and provide important baselines for comparing future, more general models.

8 Conclusion

Our work shows promise that data-limited applications may build initial systems with retrieval methods powered by recently developed embeddings. By collecting role-play transcripts and showing results in a data-limited context, we have demonstrated the potential to develop a successful simulation of a hotline visitor that novice counselors can practice with during training. We found that retrieval methods became more competitive with improved embedding methods and surpassed generative methods when more context was considered. We also found that context had impact on how difficult it was for crowdworkers to evaluate responses.

As a next step, we plan to explore better leveraging rich structure in the conversations, with a focus on the protocol that the counselors are trained to follow. There has been increased interest in blending retrieval and generation approaches by modifying prototypes retrieved from training data (Li et al., 2018; Weston et al., 2018). It is possible that such an approach would enable modifying and thus tailoring responses to similar contexts.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics*.

American Association of Suicidology. 2012. *Organization Accreditation Standards Manual*.

Allen D Andrade, Anita Bagri, Khin Zaw, Bernard A Roos, and Jorge G Ruiz. 2010. Avatar-mediated training in the delivery of bad news in a virtual world. *Journal of palliative medicine*, 13(12):1415–1419.

Ricardo Baeza-Yates, Berthier de Araújo Neto Ribeiro, et al. 2011. *Modern information retrieval*. New York: ACM Press; Harlow, England: Addison-Wesley,.

Sairam Balani and Munmun De Choudhury. 2015. Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1373–1378. ACM.

Rafael E Banchs and Haizhou Li. 2012. Iris: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, pages 37–42. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Wendi F Cross, David Seaburn, Danette Gibbs, Karen Schmeelk-Cone, Ann Marie White, and Eric D Caine. 2011. Does practice make perfect? a randomized control trial of behavioral rehearsal on suicide prevention gatekeeper skills. *The journal of primary prevention*, 32(3-4):195.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*.

Karthik Dinakar, Jackie Chen, Henry Lieberman, Rosalind Picard, and Robert Filbin. 2015. Mixed-initiative real-time topic modeling & visualization for crisis counseling. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 417–426. ACM.

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

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *ACL workshop for NLP Open Source Software*.

Madelyn S Gould, Wendi Cross, Anthony R Pisani, Jimmie Lou Munfakh, and Marjorie Kleinman. 2013. Impact of applied suicide intervention skills training on the national suicide prevention lifeline. *Suicide and Life-Threatening Behavior*, 43(6):676–691.

Madelyn S Gould, Jimmie LH Munfakh, Marjorie Kleinman, and Alison M Lake. 2012. National suicide prevention lifeline: enhancing mental health care for suicidal individuals and other people in crisis. *Suicide and Life-Threatening Behavior*, 42(1):22–35.

M Sazzad Hussain, Juchen Li, Louise A Ellis, Laura Ospina-Pinillos, Tracey A Davenport, Rafael A Calvo, and Ian B Hickie. 2015. Moderator assistant: A natural language generation-based intervention to support mental health via social media. *Journal of Technology in Human Services*, 33(4):304–329.

John Kalafat, Madelyn S Gould, Jimmie Lou Harris Munfakh, and Marjorie Kleinman. 2007. An evaluation of crisis hotline outcomes. part 1: Nonsuicidal crisis callers. *Suicide and Life-threatening behavior*, 37(3):322–337.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL*.

Ben Krause, Marco Damonte, Mihai Dobre, Daniel Duma, Joachim Fainberg, Federico Fancellu, Emmanuel Kahembwe, Jianpeng Cheng, and Bonnie Webber. 2017. Edina: Building an open domain socialbot with self-dialogues. *arXiv preprint arXiv:1709.09816*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1865–1874.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. 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. Association for Computational Linguistics.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1116–1126.

Yookoon Park, Jaemin Cho, and Gunhee Kim. 2018. A hierarchical latent structure for variational conversation modeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1792–1801.

Amber Paukert, Brian Stagner, and Kerry Hope. 2004. The assessment of active listening skills in helpline volunteers. *Stress, Trauma, and Crisis*, 7(1):61–76.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence Ann, Kathy J Goggin, and Delwyn Catley. 2017. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 1128–1137.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Rajeev Ramchand, Lisa Jaycox, Pat Ebener, Mary Lou Gilbert, Dionne Barnes-Proby, and Prodyumna Goutam. 2016. Characteristics and proximal outcomes of calls made to suicide crisis hotlines in california. *Crisis*.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, pages 583–593. Association for Computational Linguistics.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.

Aaron Smith and Dana Page. 2015. Us smartphone use in 2015. *Pew Research Center*, 1.

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. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. In *ICLR*.

Suicide Prevention Resource Center. 2007. Applied suicide intervention skills training (ASIST).

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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

Ivan Vulić, Nikola Mrkšić, Roi Reichart, Diarmuid Ó Séaghdha, Steve Young, and Anna Korhonen. 2017. Morph-fitting: Fine-tuning word vector spaces with simple language-specific rules. In *Proceedings of ACL*, pages 56–68.

Zijian Wang and David Jurgens. 2018. Its going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 33–45.

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

Jason Weston, Emily Dinan, and Alexander H Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. *arXiv preprint arXiv:1808.04776*.

A Appendices

Modified model parameters are shared below for reproducibility.

A.1 Seq2Seq Model Parameters

More information on model parameters can be found in the OpenNMT-py online documentation[2].

-dynamic_dict on

[2] http://opennmt.net/OpenNMT-py/index.html

-share_vocab on

-src_seq_length = 200

-tgt_seq_length = 200

-rnn_size = 500

-src_word_vec_size = 300

-tgt_word_vec_size = 300

-share_embeddings on

-encoder_type = brnn

-decoder_type = rnn

-rnn_type = LSTM

-layers = 2

-global_attention = general

-optim = adam

-learning_rate = 0.001

-batch_size = 4

pre-trained embedding glove.840B.300d.txt

A.2 VHCR Model Parameters

More info can be found about model parameters in the online repository[3].

-model = VHCR

-batch_size = 4

-embedding_size = 300

-encoder_hidden_size = 500

-decoder_hidden_size = 500

-context_size = 500

-z_sent_size = 50

-z_conv_size = 50

pre-trained embedding glove.840B.300d.txt

-max_sentence_length = 60

-max_conversation_length = 5

-min_vocab_frequency = 3

[3] https://github.com/ctr4si/A-Hierarchical-Latent-Structure\-for-Variational-Conversation-Modeling