End-To-End Dialogue With Sentiment Analysis Features

Alex Rinaldi[1*], Omar Oseguera[1], JoAnn Tuazon[1], and Alberto C. Cruz[1]

[1]California State University, Bakersfield, Bakersfield, CA, United States
arinaldi1@csub.edu

Abstract. Psychiatric assistance for suicide prevention does not have a wide enough reach to help the number of victims who commit suicide every year. To help people cope with suicidal thoughts when formal care is unavailable, we propose an artificial intelligence, text-based conversational agent that generates responses similar to those of a counselor. The application will offer a temporary channel for expression that serves as a transition to speaking with a professional psychiatrist. We expand upon existing approaches by utilizing sentiment analysis data, or scores that rank the emotional content of users' text input, when generating responses. We also train a response generation system based on a dataset of counseling and therapy transcripts. We posit that inclusion of sentiment analysis data provides marginally better responses based on quantitative metrics of quality. We hope our results will advance realistic conversation modeling and promote further research into its humanitarian applications.

Keywords: Sequence-to-sequence learning, dialogue system, conversational agent, chatbot, recurrent neural network, sentiment analysis

1 Introduction

Psychiatric care and counseling is a vital deterrent to suicide, but not everyone is able or willing to seek these resources. A 2011 survey by the American College Counseling Association studies this problem in the college setting. According to the survey, 80% of students who died by suicide did not seek help from their school's counseling center [1]. A 2013 Journal of American College Health survey finds at risk college students do not report suicidal thoughts primarily because they believe they don't need professional treatment, lack time, or prefer to solve personal problems without help from others [2].

One solution that overcomes these barriers is providing channels for expression that do not require formal care. Social media provides one such channel, but it does not guarantee a safe environment. Suicide hotline services can help, but have limited resources and are generally reserved for emergencies. [3] suggests methods that are completely anonymous and do not require human contact have benefits when detecting suicidal thoughts and behavior. According to [3], 2-3 times more people reported suicidal

* Corresponding author: +1-661-654-3142, arinaldi1@csub.edu

behavior when using written or computerized anonymous surveys. One possible channel to express and counter suicidal ideation while remaining anonymous and without human contact is a mobile application where users can write down their thoughts. This provides the benefit of being constantly available, but it lacks the directed nature of a survey or the responsive nature of a counseling session. In addition, mobile journaling applications like *Day One* already exist. A better solution is a journaling app that provides responses to the user's input, helping guide the expression of his or her thoughts in the same way a counselor would. Our approach takes the form of a text-based conversational agent, or an artificial intelligence that produces responses to a user's text input. The agent would be explicitly advertised as an artificial intelligence, providing the benefit of anonymity described in [3]. It would provide a temporary channel for expression that could transition into more formal psychiatric care. The rest of this paper will describe initial experiments to make this application possible.

2 Related Work and Motivation

Our approach builds upon several current approaches to developing a conversational agent, while offering three main novelties:

1. Including sentiment analysis data as part of the problem domain
2. Enhancing sequence-to-sequence dialogue generation with feature-level fusion
3. Training on a dataset of therapy and counseling sessions.

2.1 Sentiment Analysis and LIWC

Psychotherapists are required to understand and react to the emotional state of their patients. For a text-based conversational agent serving a similar purpose, some measure of the emotional content of input text is necessary for producing an appropriate response.

Detecting and quantitatively representing emotion in text is an area of research known as Sentiment Analysis, which has recently been implemented in a software application known as Linguistic Inquiry and Word Count (LIWC) [4]. LIWC provides additional features to help process the semantic and emotional content of user input. It groups words into semantic categories, including those related to basic language functions like "personal pronoun" and "auxiliary verb," as well as psychological constructs including "cognition," "positive emotional tone," and "affect" [4]. Since LIWC categories are meant to help represent the meaning of text, we posit that an artificial intelligence system should utilize LIWC categories to provide more meaningful responses.

To the best of our knowledge, [5] is the first attempt to use LIWC features to generate responses in a conversational agent application. [5] attempts to take a specific conversational objective or goal (like recommending a restaurant), and generate a response that achieves the objective while generating a user-defined "personality" (such as extroverted or introverted). LIWC scores are used to train parameters that select and com-

bine sentence structures from a handmade database to evoke the personality. This application of LIWC scores is too limited for our problem, however. Unlike in a commercial application, the conversational objective of a therapist is complex and always changing, making it difficult to model.

2.2 Retrieval and Generative Methods for Responses

[5] is an example of a retrieval-based method because it generates a response by selecting from a database of pre-defined responses (or "templates" for responses in this case). While retrieval-based methods are widely used because predefined responses are more grammatically correct and consistent, they are too limited for our application; in [5] a specific goal is required. Paper [6] attempts to produce more dynamic responses by selecting from a large dataset of responses in social media with a trained probability distribution over the possible responses, but constructing a dataset of every possible counselor response is impossible without responses sounding too contrived.

[7], [8], [9], and [10] use a more dynamic approach called sequence-to-sequence generation. Originally proposed in [11] as a method for translating between two languages, sequence-to-sequence approaches process input one word at a time and produce output one word at a time, resulting in the potential for unique responses for every input. The model is based on a combination of two recurrent neural networks (RNN). Sequence-to-sequence generation is data-driven; it is trained using a large corpus of text conversations, and attempts to learn common language patterns between all inputs and responses.

2.3 Fusion with Sentiment Analysis Features

One common problem among sequence-to-sequence implementations is that while the output follows common language patterns, it has little meaning in relation to the input. Sequence-to-sequence generators can converge to generate the most simple, frequent responses like "Yeah" as shown in Figure 2. One solution may be to include additional semantic information to help direct the learning process.

[10] proposes inputting combining additional information along with the user input. It uses a set of features representing general qualities about the speaker (such as ethnicity) in combination with each input word to help the system generate more personalized responses. This is known as feature level fusion, and is meant to improve an artificial intelligence system's performance by providing multiple sources of information. Influenced by [5], we apply this concept to our problem by combining LIWC categories for each input word with the word itself. We predict that providing the model with semantic and sentiment information about each word will help detect semantic patterns between the input and output responses and generate output that is more meaningful and emotionally relevant.

[12] and [13] provide frameworks for performing feature level fusion in RNNs for emotion recognition in video. We will use a similar approach for combining vector representations of words with LIWC category vectors as described in the technical approach. [7] and [8] both use vector embedding of words to reduce dimensionality when

working with a large vocabulary – each word in the vocabulary is assigned a lower-dimensional vector. We will therefore concatenate the LIWC categories with a vector embedding of each word.

2.4 Dialogue Corpus

While approaches [7], [9], and [10] use datasets mined from social media sites, movie and television scripts, and IT help desk conversations to train their conversational models on general dialogue, our conversational agent has a more focused purpose on providing psychological help. Therefore, we have decided to train our model on a dataset of Counseling and Therapy transcripts provided by Alexander Street [14]. The dataset includes 725 typed, anonymous transcripts of recorded therapy sessions categorized by self-reported symptoms such as suicidal ideation and depression. In total, there are approximately 86,000 input-response pairs.

3 Technical Approach

In this section we detail the sequence-to-sequence model with LIWC feature level fusion and the training process. Figure 1 provides an overview of the approach.

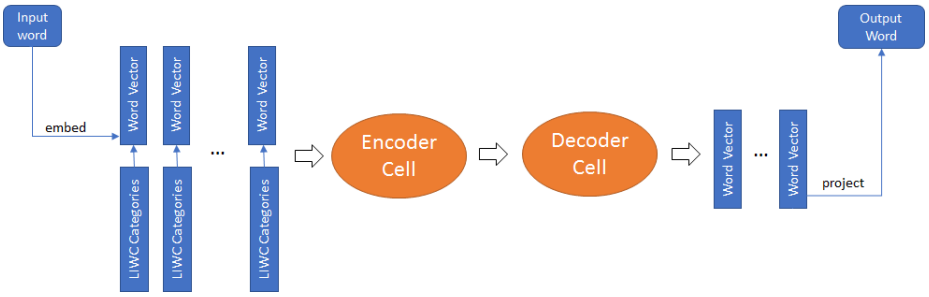

Fig. 1. Overview of proposed model.

First, all words in the dataset are replaced with an integer representing the word's position in the vocabulary for the top V_I and V_o most frequently occurring words in the client and therapist datasets, respectively. Words outside of the vocabulary are replaced with an "unknown" token, and the end of each turn in the conversation is marked with an "EOS" (end-of-sentence) token. Tokens are converted into their one-hot vector representation before being embedded and input to the encoder cell.. The client token sequences are fed to the model, and the therapist token sequences become the target for evaluating output as described below.

The model for the sequence-to-sequence RNN follows Google's provided implementation of [11], with a feature fusion method based on [12]. The encoder cell receives a word belonging to a vocabulary of size V_I embedded to a vector representation of size E. Each input word embedding $\boldsymbol{e_n} \in \mathbb{R}^E$ is concatenated with its corresponding

LIWC categories (a vector $\boldsymbol{l} \in \mathbb{R}^C$ where C is the number of categories - 1 indicates the word belongs in a category, 0 otherwise) to form the fused input vector $\boldsymbol{v_n} \in \mathbb{R}^{C+E} = [\boldsymbol{e_n}, \boldsymbol{l_n}]$.

In each encoder step, the RNN cell updates its current hidden state $\boldsymbol{h_n} \in \mathbb{R}^{C+E}$ as a function of the previous hidden state and the current fused vector:

$$h_n = f(\boldsymbol{h_{n-1}}, [\boldsymbol{e_n}, \boldsymbol{l_n}]) = f(\boldsymbol{h_{n-1}}, \boldsymbol{v_n}) \quad (1)$$

And f is a linear combination of the hidden state and input subjected to hyperbolic tangent:

$$f(\boldsymbol{h_{n-1}}, \boldsymbol{v_n}) = \tanh(\boldsymbol{W_{h,e}}\boldsymbol{h_t} + \boldsymbol{W_{v,e}}\boldsymbol{v_n}) \quad (2)$$

Where $\boldsymbol{W_{h,e}}, \boldsymbol{W_{v.e}} \in \mathbb{R}^{(C+E)\times(C+E)}$. In the output decoder, equations (1) and (2) are used to update the hidden state with different parameters $\boldsymbol{W_{h,d}}, \boldsymbol{W_{v.d}}$. The hidden state must be projected to the size of the output vocabulary V_o to achieve an output:

$$\hat{\boldsymbol{o}}_{\boldsymbol{n}} = g(\boldsymbol{h_n}) = \boldsymbol{O}\boldsymbol{h_n} \quad (3)$$

Where $\boldsymbol{O} \in \mathbb{R}^{V_o\times(C+E)}$. The output token is the index in the vocabulary as determined by the vector index with the highest value:

$$w_n = \underset{i}{\operatorname{argmax}}([\hat{o}_{n,i}]) \quad (4)$$

The parameters $\boldsymbol{W_{h,e}}, \boldsymbol{W_{v,e}}, \boldsymbol{W_{h,d}}, \boldsymbol{W_{v.d}}, \boldsymbol{O}$ are optimized through gradient descent by minimizing the log perplexity of the softmax loss between the output logit $\hat{\boldsymbol{o}}_{\boldsymbol{n}}$ and the one-hot vector representation of the target word $\boldsymbol{o_n}$ (from a therapist utterance) for all words in a given sequence of length N. Equation (5) uses the softmax function to convert the output logit into a probability distribution over the vocabulary, and (7) is the log perplexity over all probability distributions in the sequence.

$$\varsigma(\hat{\boldsymbol{o}}_{\boldsymbol{n}}) = \frac{e^{\hat{o}_{n,v}}}{\sum_{i=1}^{V_0} e^{\hat{o}_{n,i}}} \text{ for } v = 1 \dots V_o \quad (5)$$

$$loss(\hat{\boldsymbol{o}}_{\boldsymbol{n}}, \boldsymbol{o_n}) = -\sum_{n=1}^{N} \boldsymbol{o_n} \log\big(\varsigma(\hat{\boldsymbol{o}}_{\boldsymbol{n}})\big) \text{ for } n = 1 \dots N \quad (6)$$

The equations above provide a basic implementation of the RNN sequence-to-sequence model. As noted in [11], it is possible to improve the performance of the model on long sequences of words by using Long Short Term Memory (LSTM) cells while still using feature fusion. An LSTM has an input, output, and "forget" gate, so equation (3) becomes a function of the output gate in this implementation. As we will explain in the next section, the type of RNN cell is a parameter in our experiments.

4 Experimentation

In this section we discuss the methods to prepare our dataset, test it with our model, and evaluate our results.

4.1 Preparation of Data

All transcripts are processed so that every session is a single-turn conversation (the patient speaks once, then the therapist speaks once). If a client or therapist speaks for multiple turns, the turns are concatenated together. Transcriber annotations like "inaudible" are removed. All pairs of client and therapist turns from all sessions are combined into input and response datasets (without altering the order of turns in the transcripts), resulting in 86,593 client-therapist pairs.

4.2 Selection of parameters

The sizes of the input and output vocabularies V_I and V_o as well as the embedding size E are optimized through a grid search. Also, the type of RNN cell is a parameter (Basic and LSTM).

4.3 Training Specifics

Based on the implementation provided for [11], input and output sequences must have static length, so input-response pairs are categorized based on lengths of 10, 20, 30, 40, with our implementation adding 50, 60, and 70 to accommodate longer inputs. Inputs above this length are trimmed. In our dataset, 20.2% of inputs have length 70-100 and must be trimmed, resulting in 69,059 usable client-therapist pairs. Also, to help direct training target fused vectors are fed to the decoder RNN instead of generated fused vectors at each time step during training. During the test phase, output fused vectors are first projected into the vocabulary space, re-embedded, and the fused before being fed to the decoder at the next step. A variable learning rate is used, and gradient clipping is implemented to prevent exploding gradients.

4.4 Evaluation of Results

Unfortunately, there is no established metric for evaluating the quality of generated dialogue. Also, our output is subject to the constraint of being "emotionally relevant" to the user's input. [10] attempts to use human evaluation to compare models, but considering the unchecked ethical ramifications of evaluating our conversational agent and the incremental nature of advances dialogue generation (the Turing Test is far from solved), we use unsupervised methods to evaluate the quality of our responses in comparison to other models

[8] and [7] agree on using perplexity between generated responses and the target dataset as a metric for measuring dialogue generation quality. Perplexity is obtained by

taking the exponential of equation (6) and should be minimized. Our method divides the conversation dataset into training and test sets of sizes 70% and 20% respectively, and compares methods by summing the perplexities over all responses in the test set.

Influenced by [5], we also propose a simple metric to evaluate whether or not responses are emotionally relevant. Paper [5] uses LIWC scores to test whether generated sentences reflect the personality they are meant to evoke. We evaluate our model's responses against other models by root mean square difference between LIWC scores for the test set and generated responses.

Quantitative comparison of results as well as qualitative analysis will be provided for the proposed method, the baseline sequence-to-sequence implementation provided by Google for [11].

5 Discussion and Conclusion

Figure 2 demonstrates some of the shortcomings of the basic sequence-to-sequence approach implemented by Google for [11]. When trained on our Counseling and Therapy dataset, the generated responses is often "Yeah" or "Mm hm." This makes sense, as the therapist is often simply encouraging the client to continue speaking. When responding to emotional statements where the therapist needs to play a more supportive or responsive role, this kind of response may be inappropriate.

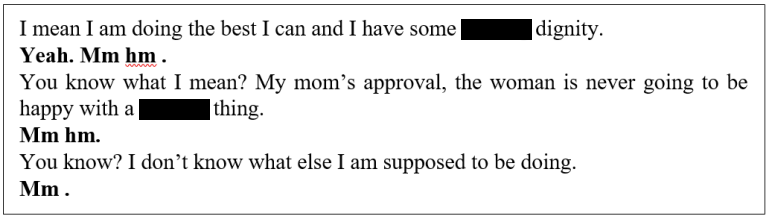

Fig. 2. Responses generated by the baseline sequence-to-sequence model are highlighted in bold. Inputs are trimmed examples from the client dataset (examples must be trimmed to accomodate the baseline sequence-to-sequence implementation). Expletives are redacted for this publication.

By including semantic and sentiment analysis in our model, we expect responses that are more meaningful in terms of the input. While qualitatively evaluating this is unfeasible in the short term, we expect our unsupervised metrics to yield marginally better results than the baseline and other state of the art methods. Future work will involve exploring methods to process longer responses and entire conversations when producing responses. We hope to encourage more research in using affective computing for humanitarian applications.

References

1. Gallagher, R.P.: National Survey of Counseling Center Directors 2011. Am. Coll. Couns. Assoc. 26 (2011).
2. Manuscript, A.: College Students at Elevated Risk for Suicide. 61, 398–406 (2014).
3. Nock, M.K., Borges, G., Bromet, E.J., Cha, C.B., Kessler, R.C., Lee, S.: Suicide and suicidal behavior. Epidemiol. Rev. 30, 133–154 (2008).
4. Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The Development and Psychometric Properties of LIWC2015. Austin, TX Univ. Texas Austin. (2015).
5. Mairesse, F.: Learning to Adapt in Dialogue Systems: Data-driven Models for Personality Recognition and Generation. 285 (2008).
6. Ji, Z., Lu, Z., Li, H.: An Information Retrieval Approach to Short Text Conversation. Arxiv - Soc. Media Intell. 1–21 (2014).
7. Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., Pineau, J.: Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. Aaai. 8 (2016).
8. Yao, K., Zweig, G., Peng, B.: Attention with Intention for a Neural Network Conversation Model. NIPS Work. Mach. Learn. Spok. Lang. Underst. Interact. 1–7 (2015).
9. Vinyals, O., Le, Q. V.: A Neural Conversational Model. ICML Deep Learn. Work. 2015. 37, (2015).
10. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A Persona-Based Neural Conversation Model. arXiv. 10 (2016).
11. Sutskever, I., Vinyals, O., Le, Q. V: Sequence to Sequence Learning with Neural Networks. Nips. 3104–3112 (2014).
12. Chao, L., Tao, J., Yang, M., Li, Y., Wen, Z.: Long Short Term Memory Recurrent Neural Network Based Multimodal Dimensional Emotion Recognition. Proc. 5th Int. Work. Audio/Visual Emot. Chall. 65–72 (2015).
13. Chao, L., Tao, J., Yang, M., Li, Y.: Multi task sequence learning for depression scale prediction from video. 2015 Int. Conf. Affect. Comput. Intell. Interact. ACII 2015. 526–531 (2015).
14. Counseling and psycotherapy transcripts series, http://alexanderstreet.com/products/counseling.

Acknowledgements

The authors (gratefully) acknowledge financial support from NSF under grant HRD-0331537 (CSU-LSAMP).