A Hierarchical Recurrent Encoder-Decoder Model for Supervised Analysis of Therapeutic Alliance in Online Therapy Conversations

Emily Tseng

Abstract

As online therapy delivered via asynchronous text-chat grows more and more popular, there emerges a need for language analysis systems that can automatically infer high-level attributes of therapy conversations. Rising tension, lessening interest, and other such attributes are easily discernable to the therapist and to any human observer, but the asynchronous nature of this type of therapy means human observers are not always on hand to make such judgments. If used wisely, automated systems might assist therapists and their supervisors in preemptively flagging conversations with the patients at highest risk for abandoning treatment, and ultimately enable them to provide better care–but to achieve such systems, we need models that can capture emerging attributes of conversations as they progress.

In this work, we provide a proof of concept for a conversational forecasting model that makes use of generative pre-training to develop a neural representation of high-level conversational dynamics. Our model uses this representation to infer the patient's score for the working alliance between patient and therapist at a given point in a therapeutic conversation. Using a proprietary dataset of transcripts and patient outcomes from a major provider of online therapy, we show that this model outperforms baselines that do not use generative pretraining. Our work extends prior literature on conversational forecasting frameworks from social media exchanges to therapy conversations, which by nature have more complex long-term dependencies. We close by identifying several opportunities for next steps.

1. Introduction

In therapeutic and caregiving contexts, a patient's felt sense of *alliance* with a care provider can make or break their treatment. This is particularly important in remote caregiving contexts like online therapy, where the vast majority of the connection between the patient and the caregiver takes place entirely over text-chat. As the popularity of this form of therapy grows, the therapists and supervisors who work on these platforms become akin to call-center employees: shuffling quickly between communication channels with an increasing and neverending caseload of patients demanding their immediate attention.

[FIXME: todo: write out the first graf of the abstract, maybe cite Verge expose]

[FIXME: add example]

In this project, we provide a proof of concept for the use of neural representations of high-level conversational dynamics in forecasting alliance between patients and therapists. We describe a model for predicting alliance at any given point in a conversation, and examine its potential using a proprietary dataset of therapy transcripts and patient-provided alliance scores from a major provider of online therapy services. We compare the performance of our model against a traditional feature-engineered method, bag-of-words, and examine variants with and without attention. Our work provides a starting point for further work on conversational forecasting in this important problem domain. To summarize, in this work we:

- develop a new model extending the current state-of-the-art in conversational forecasting to meet the challenges of online therapy conversations,
- show our model outperforms approaches that do not make use of generative pre-training prior to the supervised task, and
- lay a foundation for upcoming work examining more novel approaches using a proprietary real-world dataset of text-chat therapy transcripts.

2. Background

Online therapy data. [FIXME: What are the challenges/opportunities inherent to the data? (High dimensional, sparse, missing data, noise, structure, discrete/continuous, etc?)]

Therapeutic alliance and WAI. The psychology literature has had a longstanding interest in the role of the relationship between the therapist and the patient in successful psychotherapy. Early works posited that a beneficial attachment between patient and therapist enables the patient to trust the therapist enough to use his or her interpretations to bring about positive change (for a review, see Ardito & Rabellino 2011). Bordin 1979 defined the concept of a *working alliance* as the relational underpinning of many approaches to psychotherapy, and outlined that different types of alliance emerge from different approaches. Regardless of approach, Bordin argues, it is the *strength* of the alliance that best predicts positive outcomes.

As a way to measure alliance in a given patient-therapist relationship, Horvath & Greenberg 1989 proposed the Working Alliance Inventory (WAI), a set of self-reported scales that measure the quality of the alliance along the three dimensions defined in Bordin's theoretical framework: (1) the *bond* between patient and therapist, (2) the agreement on the *goals* of the therapy, and (3) the agreement on the *tasks* required to achieve those goals. Available in variants from the perspective of the patient, the therapist, and a third-party observer, the WAI has been validated as a reliable metric for alliance in several contexts (for a meta-analysis, see Martin et al. 2000).

3. Related Work

[FIXME: What 3-5 papers have been published in this space? How do these differ from your approach?] [FIXME: What data or methodologies do each of these works use? How do you plan to compare to these methods?]

NLP and therapy. [FIXME: cite Althoff work, CTL papers, George paper?]

Conversational forecasting. Our work adapts and extends the Conversational Recurrent Architecture for ForecasTing (CRAFT), a framework integrating generative pretraining with a supervised fine-tuning model to achieve improved predictive ability on conversation-level attributes, e.g., whether an online conversation will derail into personal attacks (Chang & Danescu-Niculescu-Mizil, 2019).

Whereas CRAFT focused on prediction of a binary outcome (derailment vs. non-derailment), our work focuses on classification of progressive segments of a conversation into one of multiple outcomes (buckets of WAI scores, as described in section 4). Chang & Danescu-Niculescu-Mizil 2019 also focus on social media conversations, including threads on Reddit and Wikipedia; in contrast, we study therapeutic conversations, which are longer and contain more long-range dependencies, thus presenting a more challenging modeling task.

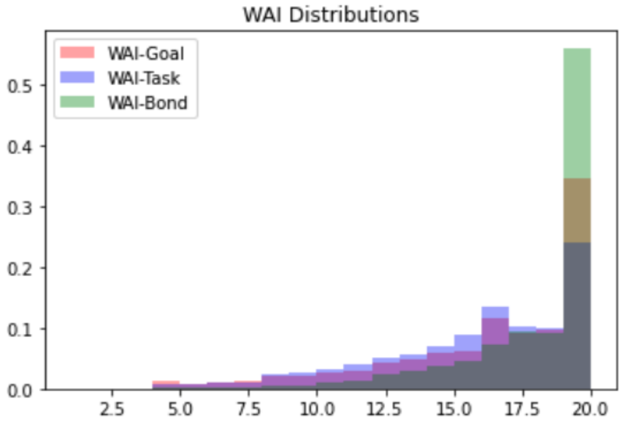

Figure 1. Distribution of WAI scores across subdimensions.

4. Dataset

We consider a dataset of therapy transcripts and associated patient outcomes from Talkspace, an online therapy platform. Due to the highly sensitive nature of therapy, we put significant effort into respecting patients' privacy and autonomy. All represented patients gave informed consent for the use of their data in research, and transcripts were anonymized by Talkspace before they were handed to our research team. Our study protocol was approved by our institutional IRB.

In total, our dataset consists of 5.7M messages exchanged between patients and therapists, representing 11,233 patients' full courses of treatment. 1,906 therapists are represented, with an average of 9 patients per therapist. Outcome annotations are provided in the form of patients' responses to surveys issued approximately every 3 weeks. In total, 13,742 WAI scores were provided by 6,702 patients. Patients provided an average of 2.1 WAI scores (range 1-24, stdev 2.2).

As depicted in Figure 1, The overall distribution of scores skewed strongly towards the positive end of the spectrum (more strongly allied). Given this, in this project we formulated our predictive task as multiclass classification between three buckets of scores: 0-14, 15-19, and 20.

5. Model

Our model is an adaptation of the Conversational Recurrent Architecture for ForecasTing (CRAFT) (Chang & Danescu-Niculescu-Mizil, 2019), which integrates a generative dialogue model and a supervised fine-tuning component to produce predictions utterance-by-utterance about some high-level conversational state.

Problem definition. We define a conversation C as a variable-length sequence of n utterances, $C =$

$\{u_1, ..., u_n\}$. Utterances are variable-length sequences of tokens w, and thus $u_n = \{w_1, ..., w_{M_n}\}$, where M_n is the length in tokens of utterance n.

Given a therapy exchange $C = \{u_1, ..., u_n\}$, we generate h_n^{con}, a neural representation of high-level conversational state up to utterance u_n. We then use h_n^{con} as the input to a classifier that predicts y_n, the label attached to the conversation up to utterance n. We define the label y_n based on the WAI score provided by the patient at utterance u_n: 0 if the score is in the bucket 0-14, 1 if 15-19, and 2 if the score is 20.

Generative component. Following Chang et al. 2019, we adopted for our generative component the hierarchical recurrent encoder-decoder (HRED) architecture proposed in Sordoni et al. 2015 and Serban et al. 2016. Built to model high-level conversational context, including temporal structure and dependencies between consecutive sequential inputs, HREDs are uniquely suited for conversational forecasting tasks.

HREDs are comprised of three component recurrent neural networks (RNNs): an utterance encoder, a conversation encoder, and a decoder. First, the *utterance encoder* generates for each utterance a semantic vector representation via its hidden state $h^{enc} \in \mathbb{R}^d_{enc}$, where d_{enc} is the desired dimension. For each token w_m in utterance n of length M, the encoder updates its h^{enc} like so:

$$h_m^{enc} \leftarrow f_{enc}^{\text{RNN}}(w_m, h_{m-1}^{enc}) \quad (1)$$

The utterance encoder's hidden state at the last step, h_M^{enc}, in theory represents an embedding for the entire utterance. Following Serban et al. 2016, h_0^{enc} is initialized as the zero vector **0**, and following Chang et al. 2019, we use the GRU (Cho et al., 2014) as our nonlinear gating function f^{RNN}.

Next, the *conversation encoder* uses the hidden states from each consecutive comment in a sequence of length N to produce an embedding h_n^{con} for the conversation up to the utterance at that point (u_N):

$$h_n^{con} \leftarrow f_{con}^{\text{RNN}}(h_{M_n}^{enc}, h_{n-1}^{con}) \quad (2)$$

The conversation encoder also initializes its hidden state h_0^{con} with the zero vector **0**, and also uses the GRU as its nonlinearity. We denote the dimension of h^{con} as d_{con}.

The *decoder* uses the embedded conversational context h_n^{con} to generate a response to utterance n. Following Sordoni et al. 2015, it does this by first initializing its own hidden state $h^{dec} \in \mathbb{R}^{d_{dec}}$ using a nonlinear activation of h_n^{con}:

$$h_0^{dec} = \tanh(Dh_n^{con} + b_0) \quad (3)$$

Where $D \in \mathbb{R}^{d_{dec} \times d_{con}}$ projects the context embedding into decoder space, and $b_0 \in \mathbb{R}^{d_{dec}}$. The decoder then updates its own hidden state for each response token using the following recurrence:

$$h_t^{dec} \leftarrow f_{dec}^{\text{RNN}}(w_{t-1}, h_{t-1}^{dec}) \quad (4)$$

The decoder then produces the next token in its response by producing a probability distribution over words from h_t^{dec}:

$$w_t = f^{out}(h_t^{dec}, w_{t-1}) \quad (5)$$

In our implementation, following CRAFT, we supplement f^{out} with attention (Luong et al., 2015), the intuition being that certain dimensions of the context encoder states c_t – output from f_{con}^{RNN} alongside h^{con} – may be more informative for decoding than others. Our final f^{out} is as follows:

$$f^{out} = \text{softmax}(\tanh(W_c[h_t^{dec}|c_t])) \quad (6)$$

Per Luong et al. 2015, we utilize the *concat*-based scoring function in our global attention to compute c_t.

Predictive component. Our predictive component uses the conversational embedding up to utterance u_n to generate a prediction for the WAI score at that utterance. We operationalize this as a multi-layer perceptron (MLP) that takes in the conversational state h_n^{con} and produces a distribution $p(Y_n|h_n^{con})$ over possible labels $Y = \{0, 1, 2\}$. Adapting from Chang et al. (2019), our MLP uses three fully-connected layers and leaky ReLU activations between each layer; but for our task of multilabel classification, we use a softmax activation. The result is a model that creates a probability distribution in which each score can be interpreted as the likelihood of the given label.

We further implemented two variants of the predictive component: classifiers with and without attention (Luong et al., 2015). The intuition here was that h_n^{con} may encode many different high-level conversational attributes, not just those that have bearing on our particular outcome variable. By learning to weight some dimensions of h_n^{con} above others, we may achieve better performance on targeted tasks.

Parameters. The parameters of our model include those used within each RNN of the generative component, $\theta_{enc}, \theta_{con}, \theta_{dec}$, as well as those used within the predictive component, θ_{clf}.

Each RNN within the generative component contains a set of parameters for its constituent GRU (see Sordoni et al. 2015 for a longer explanation). θ_{dec} also includes parameters D and b_0 transforming h_n^{con} into h_0^{dec}, as well as all the parameters for *concat*-scored global attention per Luong et al. 2015.

As for the predictive component, across both with- and without-attention variants, θ_{clf} is operationalized as a three-layer feedforward MLP. The attention variant additionally contains a layer of attention weights.

6. Training

For the generative component, training involves maximizing the log-likelihood of the provided context-reply pairs. For a given conversation C comprised of N utterances:

$$L(\theta) = \sum_{n=1}^{N} \log(P(u_n | u_{1:n-1})) \tag{7}$$
$$= \sum_{n=1}^{N} \sum_{m=1}^{M_n} \log(P(w_{n,m} | w_{n,1:m-1}, u_{1:n-1})) \tag{8}$$

Optimization is done by applying the back-propagation through time (BPTT) algorithm standard to RNN training (Rumelhart et al., 1986).

Similarly, training for the predictive component back-propagates the cross-entropy loss between the model output and the label. Note that when training the predictive component, we follow the principles of *fine-tuning* and backpropagate through the entire model, all the way back to the encoder (Howard & Ruder, 2018). This nudges the embeddings learned during pretraining towards greater applicability to the predictive task.

7. Methods

[FIXME: For each section: What are the exact details of the dataset that you used? (Number of data points / standard or non-standard / synthetic or real / exact form of the data)]

[FIXME: How did you train or run inference? (Optimization method / hyperparameter settings / amount of time ran / what did you implement versus borrow / how were baselines computed).]

[FIXME: What are the exact details of the metric used?]

Generative pre-training. We began by training our generative model following the structure outlined in section 6. From our dataset of 5.7M messages we randomly subsampled 250k pairs of contexts (sequences of utterances) and replies. Randomization allowed our model to see a variety of contexts from a variety of conversations, as opposed to repeatedly seeing subsets of the same conversations. For example, in context-reply pairs $(\{u_1, ..., u_{n-1}\}, u_n)$ and $(\{u_1, ..., u_n\}, u_{n+1})$ from the same conversation, the subsequence of utterances $\{u_1, ..., u_{n-1}\}$ repeat across both contexts.

Our generative model was implemented using Pytorch (Paszke et al., 2019), and optimized using that framework's built-in Adam optimizer. Encoder and conversation encoder learning rates were set at 0.0001, and the decoder learning rate was set at 0.0005.

Task-specific fine-tuning. All experiments with the predictive component of CRAFT-Multilabel initialized training with encoder, conversation encoder, and decoder parameters from the final generative model.

CBoW comparison. Core to the proposed applicability of the CRAFT-Multilabel framework is its ability to represent the high-level conversational dynamics that emerge from utterances in sequence (e.g., derailment, the level of working alliance present), and make predictions about those high-level conversational dynamics by using those representations as features in a downstream task.

To test whether our model was indeed representing higher-level semantics, we compared CRAFT-Multilabel against a cumulative bag-of-words (CBoW) model. Given a vocabulary V and an input context $C_n = \{u_1, ...u_n\}$, CBoW represents C_n as a $|V|$-length vector of the counts of each word in V in the context $(f(C_n))$. In our implementation, we normalized the raw counts of each word in V according to their relative importance within their contexts, quantified by their term frequency-inverse document frequency (tf-idf). CBoW is inherently suited for online learning of progressive parts of the conversation: When faced with the next utterance in the sequence, $C_{n+1} = \{u_1, ..., u_{n+1}\}$, CBoW recomputes $f(C_{n+1})$ to reflect the *cumulative* word counts across the updated context.

The $|V|$-length feature vectors $f(C)$ are used as input to a simple feedforward MLP that mimics the structure of the predictive component of CRAFT-Multilabel: that is, it produces a distribution $p(Y|f(C))$ over possible labels $Y = \{0, 1, 2\}$. Note that the CBoW by definition only captures relative word occurrences; it does not capture word order, or even boundaries between the utterances in our contexts, to say nothing of higher-level conversational dynamics. Thus, comparison of CBoW against CRAFT-Multilabel should demonstrate the effect of incorporating such conversation-level attributes into this predictive task.

Our CBoW was an MLP trained via scikit-learn's Adam optimizer (Pedregosa et al., 2011) for 200 iterations with a constant learning rate of 0.001. Similar to the predictive component of CRAFT-Multilabel, the MLP consisted of 2 hidden layers of size 100, each of which used a ReLU activation.

Data. For this proof of concept we compared [FIXME: look into train vs val splits]

8. Results

[FIXME: add no-attn variant: What were the secondary results comparing the variants of your system?]

Table 1 depicts the results of our models on our held-out test sets. As shown, CRAFT-Multilabel outperforms CBoW by roughly 8 percentage points, and both outper-

Model	Test-set Accuracy
Selecting a single class	0.3333
CBoW	0.5577
BERT	
CRAFT-Multilabel	**0.6375**

Table 1. Performance of the CBoW vs. CRAFT-Multilabel models. All trials run with a test set of 159 context/label pairs evenly balanced between the 3 classes. The random baseline (a model that only outputs 1 class) is shown for comparison.

form the naive baseline (outputting one of the three classes for every input).

9. Discussion

[FIXME: What conclusions can you draw from the results section?]

[FIXME: Is there further analysis you can do into the results of the system? Here is a good place to include visualizations, graphs, qualitative analysis of your results.]

[FIXME: What questions remain open? What did you think might work, but did not?]

10. Conclusion

In this work, we implemented, trained and tested a proof of concept for a model that can infer working alliance from the texts of a talk-therapy conversation. For a given segment of conversation, our model utilizes conversation-level embeddings learned via generative pre-training to forecast alliance scores across three buckets: 0-14, 15-19, and 20. Tested on a held-out sample of 159 context/label pairs, our model was shown to significantly outperform both a bag-of-words approach and a random baseline (Table 1).

This proof of concept opens many avenues for next steps.

- leverage BERT or GPT-2 embeddings - incorporate additional features, such as timestamps - consider variational approaches

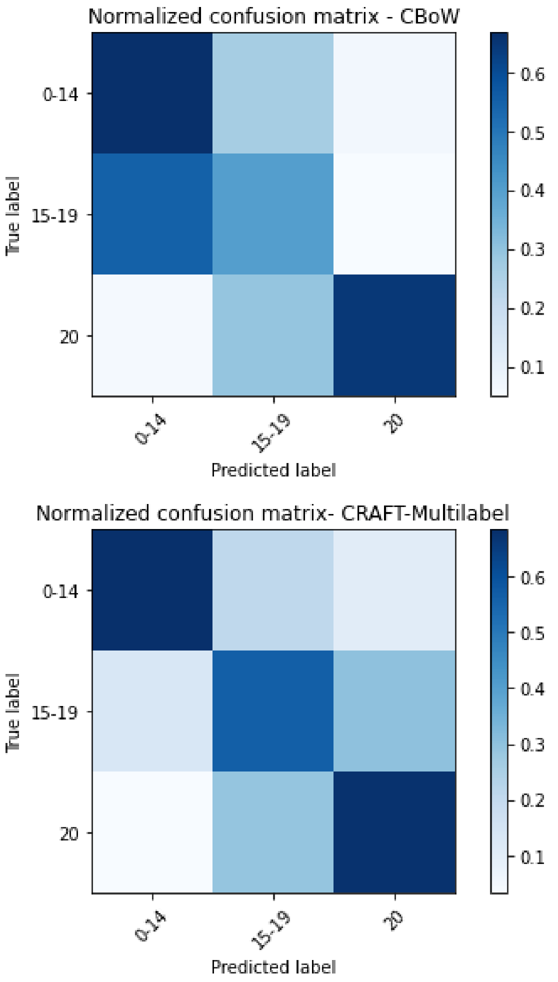

Figure 2. Confusion matrices for the CBoW vs. CRAFT-Multilabel models.

References

Ardito, Rita B and Rabellino, Daniela. Therapeutic alliance and outcome of psychotherapy: historical excursus, measurements, and prospects for research. *Frontiers in psychology*, 2:270, 2011.

Bordin, Edward S. The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, research & practice*, 16(3):252, 1979.

Chang, Jonathan P and Danescu-Niculescu-Mizil, Cristian. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Proceedings of EMNLP*, 2019.

Cho, Kyunghyun, Van Merriënboer, Bart, Gulcehre, Caglar, Bahdanau, Dzmitry, Bougares, Fethi, Schwenk, Holger, and Bengio, Yoshua. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Horvath, Adam O and Greenberg, Leslie S. Development and validation of the working alliance inventory. *Journal of counseling psychology*, 36(2):223, 1989.

Howard, Jeremy and Ruder, Sebastian. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

Luong, Minh-Thang, Pham, Hieu, and Manning, Christopher D. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

Martin, Daniel J, Garske, John P, and Davis, M Katherine. Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. *Journal of consulting and clinical psychology*, 68(3):438, 2000.

Paszke, Adam, Gross, Sam, Massa, Francisco, Lerer, Adam, Bradbury, James, Chanan, Gregory, Killeen, Trevor, Lin, Zeming, Gimelshein, Natalia, Antiga, Luca, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pp. 8024–8035, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Rumelhart, David E, Hinton, Geoffrey E, and Williams, Ronald J. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Serban, Iulian V, Sordoni, Alessandro, Bengio, Yoshua, Courville, Aaron, and Pineau, Joelle. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Sordoni, Alessandro, Bengio, Yoshua, Vahabi, Hossein, Lioma, Christina, Grue Simonsen, Jakob, and Nie, Jian-Yun. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pp. 553–562, 2015.