Deep Annotation of Therapeutic Working Alliance in Psychotherapy

Baihan Lin¹, Guillermo Cecchi², Djallel Bouneffouf²

¹Columbia University ²IBM Thomas J. Watson Research Center

baihan.lin@columbia.edu, gcecchi@us.ibm.com, djallel.bouneffouf@ibm.com

Abstract

The therapeutic working alliance is an important predictor of the outcome of the psychotherapy treatment. In practice, the working alliance is estimated from a set of scoring questionnaires in an inventory that both the patient and the therapists fill out. In this work, we propose an analytical framework of directly inferring the therapeutic working alliance from the natural language within the psychotherapy sessions in a turnlevel resolution with deep embeddings such as the Doc2Vec and SentenceBERT models. The transcript of each psychotherapy session can be transcribed and generated in real-time from the session speech recordings, and these embedded dialogues are compared with the distributed representations of the statements in the working alliance inventory. We demonstrate, in a real-world dataset with over 950 sessions of psychotherapy treatments in anxiety, depression, schizophrenia and suicidal patients, the effectiveness of this method in mapping out trajectories of patient-therapist alignment and the interpretability that can offer insights in clinical psychiatry. We believe such a framework can be provide timely feedback to the therapist regarding the quality of the conversation in interview sessions.

Index Terms: computational linguistics, computational psychiatry, natural language processing

1. Introduction

A fundamental concept in psychotherapy is the working alliance between the therapist and the patient or, more generally, the client seeking help [1]. The alliance involves several cognitive and emotional components of the relationship between these two agents, including the agreement on the goals to be achieved and the tasks to be carried out, and the bond, trust and respect to be established over the course of the therapy. Qualitative methods to quantify therapy outcomes led to the conclusion that the strength of the alliance is one of the main factors that predict success [2]. Operational methods to quantify the alliance rely of evaluative reports by patients and therapists of whole sessions, typically limited to point-scales valuation [3]. This approach does not make use of the nuances afforded by natural language, is time-consuming and difficult to follow through systematically outside of research studies; even more so is the evaluation of individual dialogue turns over the course of each session. Here we present an approach to quantify the degree of patient-therapist alliance by projecting each turn in a therapeutic session onto the representation of clinically established working alliance inventories, using language modeling to encode both turns and inventories. This allows us not only to quantify the overall degree of alliance but also to identify granular patterns its dynamics over shorter and longer time scales. We also discuss how our approach may be used as a companion tool to provide feedback to the therapist and to augment learning opportunities for training therapists.

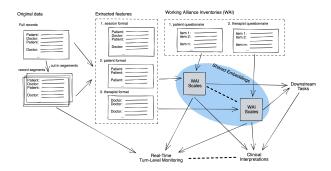


Figure 1: Analytical Framework of working alliance analysis

2. Problem Setting

2.1. Working alliance analysis

Algorithm 1 Working Alliance Analysis (WAA)

```
1: for \mathbf{i} = 1, 2, \cdots, T do
2: Automatically transcribe dialogue turn pairs (S_i^p, S_i^t)
3: for (I_j^p, I_j^t) \in \text{inventories } (I^p, I^t) do
4: Score W_j^{p_i} = \text{similarity}(Emb(I_j^p), Emb(S_i^p))
5: Score W_j^{t_i} = \text{similarity}(Emb(I_j^t), Emb(S_i^t))
6: end for
7: end for
```

The figure above is an outline of the analytic framework. We take the full records of a patient, or a cohort of patients belonging to the same condition. We either use it as is before the feature extraction, or we truncate them into segments based on timestamps or topic turns. As you can see, the original format is in pairs of dialogues. We can extract the features in three ways: first, we can use the full pairs of dialogues; second, we can only extract what the patient says; or the third option, we only extract what the doctor says. The three feature formats all have their pros and cons. The dialogue format contains all information, but the intents within the sentences come from two individuals, so they might mix together. The patient format contains the full narrative of the patients, which is usually more coherent, but it's only part of the story. The therapist format, which people in computational psychiatry also believes to be some kind of semantic labels of what the patient feels, can be informative, but they can also be sometimes too simplistic.

When we have the features, we compare the working alliance inventories with the embeddings. Algorithm 1 outlines the process. During the session, the dialogue between the patient and therapist are transcribed into pairs of turns (such as the example in Figure 2). We denote each patient response turn as S_i^p followed by a therapist response turn S_i^t . They are treated as

```
PATIENT: I don't know. I was laughing at myself for thinking it. 
THERAPISI: Yeah. 
PATIENT: Um, like I don't know, was I being a martyr? I don't know. 
THERAPIST: No it sort of sounds like you feel like you're doing me a favor or 
something? 
PATIENT: Um, I don't know. 
THERAPIST: Like he's getting more out of this than I am. [00:03:07] 
PATIENT: No, I don't think that. Naw, I'm just cranky. 
THERAPIST: Yeah I'm hearing that, but I'm trying to explain what the, what 
that's about. I mean you ought to be glad I'm here at all.
```

Figure 2: Example dialogue from psychotherapy transcripts

a dialogue pair. The inventories of working alliance questionnaires also come in pairs: I^p for the patient (or client), and I^t for the therapist. They each consist of 36 statements. We embed both the dialogue turns and the inventories with deep sentence or paragraph embeddings, and then compute the cosine similarity between the embedding vectors of the turn and its corresponding inventory vectors. With that, for each turn (either by patient or by therapist), we obtain a 36-dimension working alliance score. We will describe in section 3.1 the specific scales of our inferred working alliance scores which introduces interpretable information into our framework.

Here are a few downstream tasks and user scenarios that can plugged to our analytical frameworks. We can either use these extracted weighted topics to inform whether the therapy is going the right direction, whether the patient is going into certain bad mental state, or whether the therapist should adjust his or her treatment strategies. This can be built as an intelligent AI assistant to remind the therapist of such things.

2.2. Psychotherapy Transcript Dataset

The Alex Street Counseling and Psychotherapy Transcripts dataset consists of transcribed recordings of over 950 therapy sessions between multiple anonymized therapists and patients. This multi-part collection includes speech-translated transcripts of the recordings from real therapy sessions, 40,000 pages of client narratives, and 25,000 pages of reference works. These sessions belong to four types of psychiatric conditions: anxiety, depression, schizophrenia and suicidal. Each patient response turn S_i^p followed by a therapist response turn S_i^t is treated as a dialogue pair. In total, these materials include over 200,000 turns together for the patient and therapist and provide access to the broadest range of clients for our linguistic analysis of the therapeutic process of psychotherapy.

3. Methods

3.1. Working Alliance Inventories

The Working Alliance Inventory (WAI) is a set of self-report measurement questionnaire that quantifies the therapeutic bond, task agreement, and goal agreement [3, 4, 5]. Since the original 12-item version [4], the inventory has used parallel versions for clients and therapist with good psychometric properties and helped establish the importance of therapeutic alliance in predicting treatment outcomes. The modern version of the inventory consists of 36 questions (Figure 3), and the participant is asked to rate each item on a 7-point scale (1=never, 7=always)[5]. The WAI aims to (1) measure alliance factors across all types of therapy, (2) document the relationship between the alliance measure and the corresponding theoretical constructs underlying the measure, and (3) related the alliance measure to a unified theory of therapeutic change [6].

I felt uncomfortable with
_ and I agreed about the things I will need to do in therapy to help improve my situation.
I was worried about the outcome of the sessions.
What I was doing in therapy gave me new ways of looking at my problem.
_ and I understood each other.
_ perceived accurately what my goals were.
I find what I was doing in therapy confusing.
I believe _ liked me.
I wish _ and I could have clarified the purpose of our sessions.
I disagreed with about what I ought to get out of therapy.
I believe the time and I were spending together was not spent efficiently.
did not understand what I was trying to accomplish in therapy.
I was clear on what my responsibilities were in therapy.
The goals of the sessions were important for me.

Figure 3: Example statements in working alliance inventory

TASK scale:	2,	4,	7,	11,	13,	15,	16,	18,	24,	31,	33,	35
Polarity	+	+	-	-	+	-	+	+	+	-	-	+
BOND scale:	1,	5,	8,	17,	19,	20,	21,	23,	26,	28,	29,	36
Polarity	-	+	+	+	+	-	+	+	+	+	-	+
Polarity GOAL scale:	3,				+ 12,				+ 27,	•		+ 34

Figure 4: Keys to the three scales of working alliance inventory

Operationally, the goal is to derive from these 36 items three alliance scales: the task scale, the bond scale and the goal scale. They measures the three major themes of psychotherapy outcomes: (1) the collaborative nature of the patient-therapist relationship; (2) the affective bond between therapist and patient, and (3) the therapist's and patient's capabilities to agree on treatment-related short-term tasks and long-term goals. The score corresponding to the three scales comes from a key table (Figure 4) which specifies the positivity or the sign weight to be applied on the questionnaire answer when summing in the end. The full scale is simply the sum of the scores of the three scales. The key table is like a weighting matrix that specifies the directionalities of the scales.

3.2. Sentence Embeddings

In principle, any sentence or paragraph embeddings can help us characterize the dialogue turns and inventories. In this work, we used two deep embeddings. The Doc2Vec embedding [7] is a popular unsupervised learning model that learns vector representations of sentences and text documents. It improves upon the traditional bag-of-words representation by utilizing a distributed memory that remembers what is missing from the current context. The other embedding we evaluated is the SentenceBERT [8], which modifies a pretrained BERT network by using siamese and triplet network structures to infer semantically meaningful sentence embeddings. With these two deep embeddings, we embed the turn-level entries (either the dia-

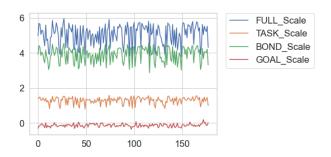


Figure 5: Example trajectory of the working alliance scores

¹https://alexanderstreet.com/products/counseling-andpsychotherapy-transcripts-series

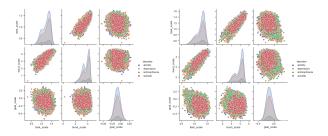


Figure 6: Relational plots of the working alliance score scales (left: patient version; right: therapist version)

Alliance score ranges in four scales

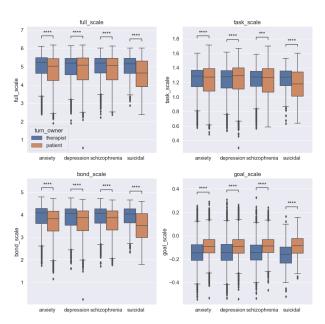


Figure 7: Box plots of the working alliance scores

logue turn in the transcripts, or the statement item in the working alliance inventories) into vectors of 300 or 384 dimensions. And then compute the cosine similarity between the vector at certain turn and an inventory entry.²

4. Results

Figure 5 is an example time series of an anxiety psychotherapy session. We see that the alliance scores varies across the scales. If we investigate the relationship among the scales, we observe that the task scale positively correlates with the bond scale in both versions, while the goal scale slightly negatively correlates with the task scale in the therapist version (Figure 6).

We investigate the consistency of the alliance estimation by the patient vs. the therapist. Overall, comparing to the patient estimates, we observe that the therapist tends to overestimate the working alliance. More specifically, the therapists overestimates the task and bond scales, but underestimates the goal scale. These differences are all statistically significant (p < 0.001).

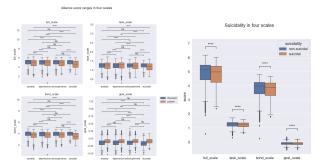


Figure 8: Alliance scores across disorders

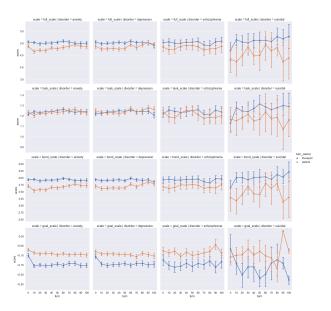


Figure 9: Four-way ANOVA of the alliance dynamics

Between the disorders, the alliance scores between anxiety and depression, and between anxiety and schizophrenia, are all significantly different in both the therapist and patient versions (p < 0.001). As in Figure 8, the suicidality can be significantly detected based on the working alliance scores of all four scales.

We also perform a four-way ANOVA upon the alliance scores as time-series sequences. Figure 9 demonstrates the difference of the dynamics of the therapeutic alliance across the psychiatric conditions. We observe that they vary by both the disorders and scales, and there appears to be certain trends along the temporal dimension (x-axis in each subplot). This is further supported by the linear regression analysis (Figure 10) that the patients with anxiety and depression have an upward alliance rating while their therapists tend to believe otherwise, and the therapists of the suicidal patients tend to have a higher alliance

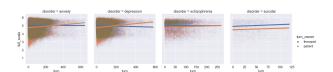


Figure 10: Regression analysis of the temporal progression of the working alliance score

²Given the space limit, the results for the Doc2Vec are shown later, while the SentenceBERT results will be included in the extended online version as supplementary materials.

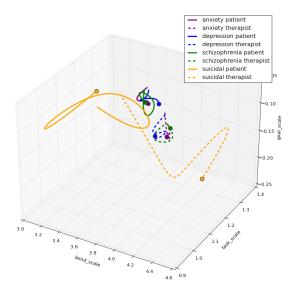


Figure 11: The average 3d trajectories of different classes of psychiatric conditions in the alliance space (the dot meaning the end points of the trajectories)

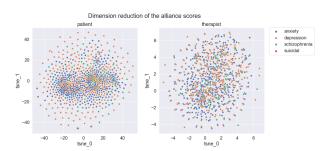


Figure 12: Dimension reduction of the alliance trajectories

rating than their patients.

We can also map out their trajectories in the alliance space of the three major scales (task, bond and goal). As in Figure 11, we plot the average trajectories of different psychiatric conditions and notice that the suicidal trajectories are much more spread out in the bond and task scales (which aligns with the findings in the ANOVA plots). Based on the directionality, the suicidality trace shows a significant divergence trend. This is the first step of a potential turn-level resolution temporal analysis of the working alliance. We can be generalize in a sense that with this approach one can go over your sessions (as a therapist) and analyze the dynamics afterwards.

Given these time-series, we can visualize them with dimension reduction techniques such as t-SNE. Because the psychotherapy sessions come in different lengths, we compute the dynamic time wrapping distances between the session trajectories of 36-dimension alliance scores, and then use this pairwise distance matrix to perform the t-SNE unsupervised learning. Figure 12 presents the difference between the manifolds of the therapist alliance trajectory space and the patient alliance space. We notice that the patient trajectories have two major clusters of alliance, while the therapist only has one. This is consistent with what we observed in the relational plots 6 that the patient alliance scores in the task and bond scales follow a bi-modal distribution.

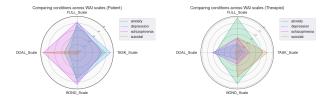


Figure 13: Radar plots of the working alliance scores

We can further aggregate the alliance score by averaging all time points together all into the four scales. To plot the scale with respect to one another in a single plot, we normalize each scale to standard normal and present the radar plots of the scalewise features of the patient and the therapist (Figure 13). We observe that the suicidal patient are comparatively most imbalanced, large only in the goal scale and small in all others. While, on the therapist version, it is the opposite, which aligns with the observation made in the 3d trajectories.

5. Discussion

Our analytic approach reveals several insightful features of the therapeutic relationship. We observe systematic differences in the mean inferred alliance scores between patients and therapists, and also across disorders. However the in-session evolution of the inferred scores provide a much more interesting perspective. In particular, while all conditions show a systematic misalignment of scores between patients and therapists, this is significantly starker for suicidality, something that can be observed in the mean as well as in the time trace for full and sub-scales. In contrast, anxiety and depression display a clear trend for the full and the bond scales to converge as the sessions progress, something not present in the task and goal scales, nor in schizophrenia or suicidality. These features of the therapeutic dialogue can be mapped to what in psychiatry is usually called alignment and plays an important symptomatic and diagnostic role in several neuropsychiatric conditions, e.g., in relation to the hypothesis of Theory of Mind for schizophrenia [9]. By analyzing past sessions, and eventually sessions in real time, trained therapists may be able to identify key segments of the therapy leading to breakthroughs, compounding their expertise with further causal/predictive analytic modeling, while trainees may sharpen their intuition by reading or watching annotated versions of sessions conducted by experts. Needless to say, coupled with a generative language model and further statistical optimization, it may be possible to design limited chatbots to engage patients in triage and emergency response [10].

6. Conclusions

We have presented an approach that combines the state-of-theart language modeling with the knowledge and practical expertise in psychotherapy, as captured in therapy-evaluation inventories, to provide a uniquely granular representation of the evolution of the interaction of patients and therapists. While here we focus specifically on the Working Alliance Inventory, our method is generic and can be extended to the broader spectrum of assessment instruments. Finally, it would be possible to refine and further validate the language-based estimation of working alliance by providing punctuated rater evaluations as inference anchors. Next steps include predicting these inference anchors as states (like [11, 12]) and training chatbots as reinforcement learning agents given these states (like [13, 14, 15]).

7. References

- [1] E. S. Bordin, "The generalizability of the psychoanalytic concept of the working alliance." *Psychotherapy: Theory, research & practice*, vol. 16, no. 3, p. 252, 1979.
- [2] B. E. Wampold, "How important are the common factors in psychotherapy? an update," World Psychiatry, vol. 14, no. 3, pp. 270– 277, 2015.
- [3] A. O. Horvath, "An exploratory study of the working alliance: Its measurement and relationship to therapy outcome," Ph.D. dissertation, University of British Columbia, 1981.
- [4] T. J. Tracey and A. M. Kokotovic, "Factor structure of the working alliance inventory." *Psychological Assessment: A journal of consulting and clinical psychology*, vol. 1, no. 3, p. 207, 1989.
- [5] D. J. Martin, J. P. Garske, and M. K. Davis, "Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review." *Journal of consulting and clinical psychology*, vol. 68, no. 3, p. 438, 2000.
- [6] A. O. Horvath and L. S. Greenberg, The working alliance: Theory, research, and practice. John Wiley & Sons, 1994, vol. 173.
- [7] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*. PMLR, 2014, pp. 1188–1196.
- [8] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," arXiv preprint arXiv:1908.10084, 2019.
- [9] J. N. De Boer, S. G. Brederoo, A. E. Voppel, and I. E. Sommer, "Anomalies in language as a biomarker for schizophrenia," *Current opinion in psychiatry*, vol. 33, no. 3, pp. 212–218, 2020.
- [10] S. Garg, I. Rish, G. Cecchi, P. Goyal, S. Ghazarian, S. Gao, G. Ver Steeg, and A. Galstyan, "Modeling dialogues with hash-code representations: A nonparametric approach," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3970–3979.
- [11] B. Lin, D. Bouneffouf, G. Cecchi, and R. Tejwani, "Neural topic modeling of psychotherapy sessions," *arXiv preprint*, 2022.
- [12] B. Lin, D. Bouneffouf, and G. Cecchi, "Predicting human decision making in psychological tasks with recurrent neural networks," arXiv preprint arXiv:2010.11413, 2020.
- [13] B. Lin, G. Cecchi, D. Bouneffouf, J. Reinen, and I. Rish, "A story of two streams: Reinforcement learning models from human behavior and neuropsychiatry," in *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent* Systems, 2020, pp. 744–752.
- [14] ——, "Models of human behavioral agents in bandits, contextual bandits and rl," in *International Workshop on Human Brain and Artificial Intelligence*. Springer, 2021, pp. 14–33.
- [15] ——, "Unified models of human behavioral agents in bandits, contextual bandits and rl," arXiv preprint arXiv:2005.04544, 2020.