Interspeech 2018
2-6 September 2018, Hyderabad, India

Language Features for Automated Evaluation of Cognitive Behavior Psychotherapy Sessions

Nikolaos Flemotomos[1], *Victor R. Martinez*[2], *James Gibson*[1], *David C. Atkins*[3], *Torrey A. Creed*[4], *Shrikanth Narayanan*[1,2]

[1]Department of Electrical Engineering and [2] Department of Computer Science,
University of Southern California, Los Angeles, CA, USA
[3] Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA
[4] Department of Psychiatry, University of Pennsylvania, Philadelphia, PA, USA

flemotom@usc.edu, victorrm@usc.edu, jjgibson@usc.edu, datkins@u.washington.edu,
tcreed@pennmedicine.upenn.edu, shri@sipi.usc.edu

Abstract

Cognitive Behavior Therapy (CBT) is a psychotherapy treatment that uses cognitive change strategies to address mental health problems. Quality assessment of a CBT session is traditionally addressed by human raters who evaluate recorded sessions along specific behavioral codes, a cost prohibitive and time consuming method. In this work we examine how linguistic features can be effectively used to develop an automatic competency rating tool for CBT. We explore both standard, widely-used lexical features and domain-specific ones, adapting methods which have been successfully used in similar psychotherapy session coding tasks. Experiments are conducted on manual transcripts of CBT sessions and on automatically derived ones, thus introducing an end-to-end approach. Our results suggest that a real-world system could be developed to automatically evaluate CBT sessions to assist training, supervision, or quality assurance of services.

Index Terms: cognitive behavior therapy, language features, behavioral signal processing

1. Introduction

As interventions based on spoken language, the necessary information for assessing psychotherapy quality is encoded in therapists' and patients' speech and language characteristics. Thus, the research-based method for measuring the competence of mental health providers is to use recorded sessions, which are rated by human coders. However, the time and cost barriers introduced by such a method lead to poor feasibility in real-world settings [1], which has raised a growing interest in computational approaches to psychotherapy quality assessment during the last few years [2].

Common natural language processing techniques, such as n-gram based methods [3] and topic models [4] have been successfully applied for the classification of Motivational Interviewing (MI) sessions, a psychotherapy approach used for treatment of conditions such as addiction. Aiming at better capturing the psychometric properties encoded in therapist's language or the sequential and dyadic interaction between the therapist and the patient, n-gram features have been combined with features inspired by psycholinguistic norms [5] and by dialog act tagging [6]. Non-lexical speech characteristics, such as speech rate [7] and prosodic features [8] have also been studied. Although most of the research efforts focus on the therapist's language and speech, it has been shown that examining patient's language can be beneficial for specific behavior cues [9]. More recently, text-only approaches to assess MI have been possible thanks to deep learning models [10–13].

Motivated by the line of work in the MI domain, we explore ways to automatically classify Cognitive Behavior Therapy (CBT) sessions. CBT is an evidence-based psychotherapy, with strong research support across a range of mental health problems [14]. However, to ensure high quality in real-world settings, it is vital to have performance-based measures of the providers' competency [15]. Although such measures exist [16], the actual rating has traditionally been done by human coders and has mainly been used in research studies.

The differences between the MI and CBT domains are twofold. First, the topics discussed during a CBT session are not restricted to substance use as in MI. Second, the MI codes are almost entirely focused on the psychotherapy session as a process by e.g. rating the different types of questions and reflections [17, 18], while the CBT coding system deals both with the process and the particular content of the session (e.g. was homework assigned?). As such, a challenge arises on finding features that can capture those two different angles of the psychotherapy session.

The current work is, to the best of our knowledge, the first effort for automated CBT evaluation; a step towards scaling up CBT quality assessment to real-world use. To that goal, we examine how different sets of linguistic features can be used to classify high vs. low quality CBT sessions across a range of individual quality metrics, as well as a total quality score, hoping that this will trigger further research in applying computational methods to CBT quality assessment.

2. Cognitive Behavior Therapy

CBT is a short-term psychotherapy teaching patients skills for creating shifts in their patterns of thinking and responding to situations. It is based on the cognitive model, according to which the link between a person's thoughts and feelings is of crucial importance and a primary factor contributing to psychological problems and mental illness [14]. CBT was originally developed with a focus on depression [19], but over the years it has expanded and adapted to a variety of problems. It is a treatment customized to the individual patient, where the therapist works towards the modification of the patient's belief system in a way that will lead to long-lasting behavioral changes.

The gold-standard measure for CBT quality is the Cognitive Therapy Rating Scale (CTRS) [16]. Each of the 11 session-level codes listed in Table 1 is scored on a 7-point Likert scale

10.21437/Interspeech.2018-1518

that ranges from 0 (poor) to 6 (excellent) [14]. A competent delivery of CBT is represented by a total CTRS score [tot], which is the sum of all the codes, greater than or equal to 40. Although all the codes play an equally important role for the evaluation of the session, it seems they can be naturally divided into a few broad categories. Thus, some of them are associated with the management and structure of the session (setting a satisfactory agenda [ag], getting feedback from the patient [fb], using time efficiently [pt], assigning cognitive therapy homework [hw]), some of them are based on the establishment of a good relationship with the patient (displaying a degree of warmth and concern [ip], setting up collaboration [co], having a good ability to empathize [un]), and finally some of them are related to a more abstract conceptualization (helping patient to see new perspectives [gd], focusing on key thoughts and behaviors [cb], having a consistent strategy for change [sc], applying cognitive-behavioral techniques [at]). For details on CTRS and its properties the reader may refer to [16] and [20].

Table 1: *The 11 CBT quality codes defined by CTRS.*

abbreviation	meaning
ag	agenda
fb	feedback
un	understanding
ip	interpersonal effectiveness
co	collaboration
pt	pacing and efficient use of time
gd	guided discovery
cb	focusing on key cognitions and behaviors
sc	strategy for change
at	application of cognitive-behavioral techniques
hw	homework

3. Method

In this work we are interested in the binary classification problem to distinguish whether CBT delivery is satisfactory or in need of improvement, which in a practical context could suggest a need for additional training or alternative strategies for a particular patient. We address the classification with respect to the total CTRS as well as with respect to each individual CTRS code, binarizing the available data as described in Section 4. The classifier used is always a linear Support Vector Machine, where the samples are weighted inversely proportionally to their class frequencies. The usage of a more sophisticated classifier would probably lead to better results, but we are mainly interested in investigating the effect of the various linguistic features.

3.1. Session Decoding Pipeline

By applying a text-based CBT evaluation, we may avoid the burden of manual behavioral coding, but we introduce the burden of manual transcription. As an alternative, we propose an end-to-end evaluation, using as input the raw audio data which, after preprocessing, is given to a voice activity detection, diarization, role matching, and Automatic Speech Recognition (ASR) system, following the pipeline described in [21]. However, the role matching module proposed in [21] is biased towards the therapist because the authors are primarily interested in the therapist's linguistic patterns. Thus, in this work we are using a slightly different approach for that module.

Let's assume that after diarization we have a set of utterances S_1 labeled as belonging to 'speaker 1' and another set S_2 for 'speaker 2'. Assuming we do have a subset of manually transcribed sessions, we construct a Language Model (LM) for the therapist T (L_T) and a LM for the patient P (L_P). We, then, assign S_i to either T or P by estimating the corresponding perplexities $pp(\cdot)$ and using the following criterion:

- if $pp(S_i|L_T) < pp(S_i|L_P)$ and $pp(S_j|L_P) < pp(S_j|L_T)$ assign S_i to T, S_j to P, where $i, j \in \{1, 2\}$
- else $k \triangleq \underset{i\in\{1,2\}}{\text{argmax}}|pp(S_i|L_T) - pp(S_i|L_P)|$
  - if $pp(S_k|L_T) < pp(S_k|L_P)$ assign S_k to T, S_m to P
  - else assign S_k to P, S_m to T

  where $m \in \{1, 2\} \setminus \{k\}$

3.2. Feature Extraction

Maybe the most widely used feature set for document representation is the set of occurrences or frequencies of n-grams in the document. In this work we weight each n-gram by the term frequency - inverse document frequency (tf-idf) [22]. We only consider unigrams, since experimentation showed that higher-order n-grams do not lead to improvements.

Semantic vector space representation of words, known as word embeddings, is also commonly used for document classification [23]. Here we use the Global Vectors for Word Representation (GloVe) [24], pretrained on 840B tokens found in the Web, which are 300-dimensional features representing each word. A session embedding is the mean of the session's utterance embeddings, which are the means of the utterances' word embeddings.

Inspired by the promising results in [5], we are also experimenting with the Linguistic Inquiry and Word Count (LIWC) features [25], as well as with Psycholinguistic Norm Features (PNFs) [26]. The former count occurrences of words belonging to particular categories, according to pre-defined category dictionaries. Keeping only the 'psychological processes' and the 'personal concerns' dimensions gives us a 46-dimensional vector for each session. The latter are lexical norms encoding aspects such as emotion or age. Considering the 13 dimensions described in [5] associated with 3 part of speech tags (adjectives, adverbs, verbs) leads to a 39-dimensional vector for each word. The mean is computed for each utterance and the mean of means is used as a session representation.

In order to capture some form of dyadic interaction, we label each utterance with one Dialogue Act (DA) from the categories 'question', 'statement', 'agreement', 'appreciation', 'incomplete', 'backchannel', 'other'. Those are predicted by a linear chain Conditional Random Field (CRF) model pretrained on the Switchboard DAMSL dataset [6]. The training is done using as observations the trigrams and the speaker labels within a local context window with size 0 and 1 for the two sets of observations, respectively. The features finally extracted are the total number of each DA encountered within a session.

Apart from the word embeddings, all the features explored are highly interpretable, a crucial issue for the task in hand, in case the final goal is not just to classify, but to give feedback for improvement to the clinician.

4. Datasets

The Beck Community Initiative partners provide high-quality training in CBT to community clinics and, through this work, have generated an archive of over 5000 recorded CBT sessions, nearly 2000 of which have been manually coded with the

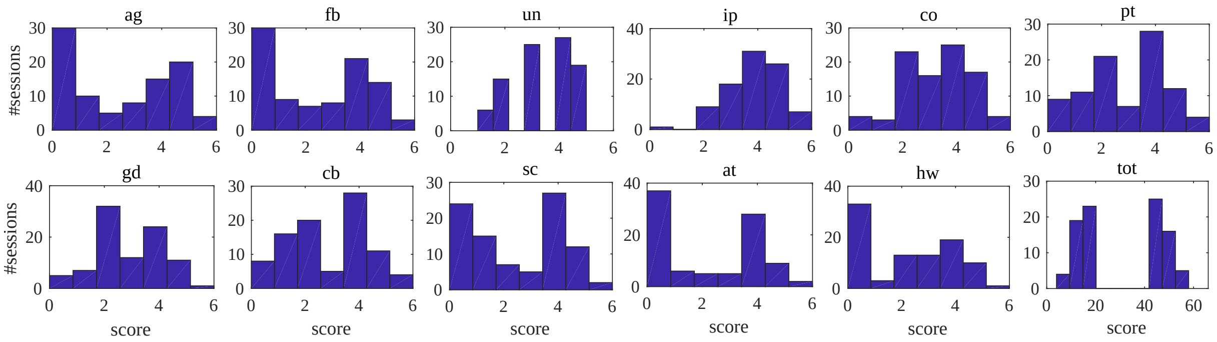
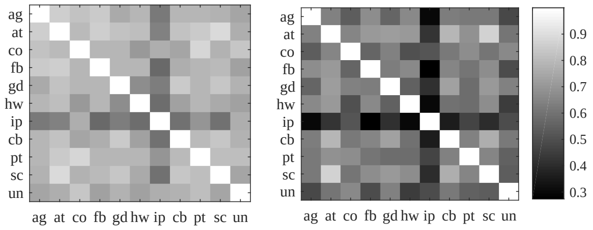


Figure 1: *Distribution of the CTRS codes for the* trans *set.*

CTRS [15]. For this work we considered a subset of 386 outpatient sessions with adult patients from 131 therapists where the CTRS codes are available (*adout* set).

After downsampling the audio files to 16kHz and keeping one audio channel, we estimated the mean Signal-to-Noise Ratio (SNR) and, out of those with a mean SNR greater than 7dB, we sent for manual transcription the 50 sessions with the highest and the 50 with the lowest total CTRS. Some of them went missing due to formatting issues, so we finally kept 92 sessions from 70 therapists for further experimentation (*trans* set).

The distribution of the codes for the *trans* set is presented in Figure 1. As observed, most of the codes tend to be concentrated towards the extremes of the scale, which is led by our sampling strategy of the sessions chosen to be sent for transcription and from the high correlation between the independent codes of the sessions finally chosen (Figure 2a). However, a high positive correlation between the codes is observed not only in the *trans*, but also in the entire *adout* set (Figure 2b). That indicates that if a therapist is considered to be "good", it is likely that he did a good job at every CBT-related aspect and is a sign that we could instead focus on predicting a subset of the codes, an issue that needs further investigation.



(a) trans *set*

(b) adout *set*

Figure 2: *Correlation matrices of the CTRS codes.*

We binarized all the CTRS codes, labeling as negative the sessions with a code less than 3, since sessions with a CTRS code greater or equal to 3 are labeled as 'satisfactory' according to the CTRS coding manual [14]. In the case of the total CTRS, the corresponding threshold score was 40, as done in clinical trials. The ratio of positive to negative samples for each code is given in Figure 3.

5. Experiments and Results

We extracted the feature sets described in Section 3.2 for the therapist (T) and the patient (P) and classified the sessions in the *trans* set with respect to each CTRS code using features independently, with the results being reported in Table 2, while our choice to use the tf-idf tranformation over simple unigrams is justified by their comparison in Table 3. The baseline classifier is one which always chooses the majority class. The results are based on a 5-fold cross-validation across therapists, so that same therapists do not appear in both the training and test folds. All the features were standardized, except for the tf-idfs which were l_2-normalized. To reduce the dimensionality of tf-idfs, we selected the K best features based on a univariate F-test. After cross-validation on the total CTRS, K was chosen to be 38 and 34 for T and P, respectively. As observed, the features corresponding to T did almost always a better job than the ones corresponding to P. In particular, tf-idf_T almost consistently yielded the best performance, but this was not the case for the codes `ip` and `co`, while the results were also poor for `un`. Those three are at the same time the most skewed codes (Figure 3) and the ones related to the therapist-patient relationship and the empathic skills of the former. It seems that simple linguistic patterns fail to reveal such complex characteristics. The PNFs, the DAs and the GloVe embeddings also demonstrate good predictive power, although the corresponding results are substantially worse than the tf-idfs (except for the cases of `ip`, `co` and `un`).

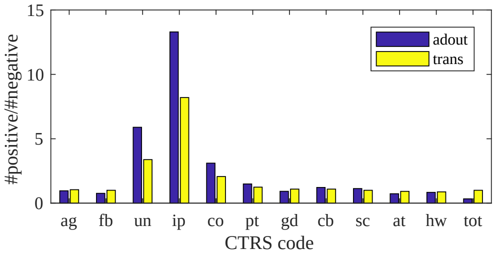


Figure 3: *Ratio of positive to negative data samples.*

We also experimented with fusion schemes of the classifiers, either at the feature or at the decision level with majority voting or stacking, but we found no substantial improvements.

In order to investigate which are the most informative words leading to such a good performance of the tf-idfs, we did two independent tests. First, for each fold we performed a backward selection to find the subset of the 5 best features/words. The words 'homework', 'agenda', and 'evidence' were consistently among the best five words at every fold, both for the total CTRS and for most of the codes. Second, we computed the correlation of all the words in the feature set with the codes to be predicted. Again, those three words yielded the highest correlation values with the Spearman correlation being in the range $[0.72, 0.81]$ for the total CTRS and the corresponding p-values being < 0.001. The importance of those specific words was expected, as far as the `hw`, `ag` and probably `gd` (where the therapist has to help patient see new perspectives by examining evidence [14]) codes are concerned, but we were surprised that they played such a substantial role for all the codes and the to-

Table 2: *Averaged F_1 score for the binary classification of CBT sessions on the* trans *set using the independent feature sets proposed.*

	tf-idf_T	pnf_T	liwc_T	glove_T	da_T
ag	**0.91**	0.69	0.45	0.82	0.78
fb	**0.83**	0.69	0.48	0.82	0.75
un	**0.55**	0.47	0.46	0.51	0.52
ip	0.46	0.43	0.41	**0.62**	0.46
co	0.63	0.56	0.49	**0.65**	0.57
pt	**0.87**	0.63	0.51	0.77	0.70
gd	**0.85**	0.67	0.47	0.74	0.71
cb	**0.85**	0.70	0.52	0.76	0.75
sc	**0.86**	0.69	0.50	0.81	0.78
at	**0.86**	0.71	0.50	0.76	0.75
hw	**0.82**	0.61	0.49	0.71	0.70
tot	**0.86**	0.71	0.49	0.81	0.76

	tf-idf_P	pnf_P	liwc_P	glove_P	da_P	baseline
ag	0.61	0.73	0.35	0.78	0.68	0.32
fb	0.62	0.69	0.32	0.73	0.67	0.32
un	0.45	0.48	0.38	0.47	0.51	0.43
ip	0.56	0.44	0.39	0.47	0.49	0.57
co	0.57	0.61	0.33	0.71	0.57	0.40
pt	0.65	0.64	0.38	0.68	0.60	0.35
gd	0.54	0.66	0.41	0.64	0.64	0.34
cb	0.57	0.64	0.35	0.59	0.62	0.32
sc	0.58	0.68	0.38	0.69	0.61	0.31
at	0.67	0.63	0.38	0.70	0.61	0.34
hw	0.56	0.66	0.40	0.70	0.67	0.34
tot	0.63	0.68	0.37	0.71	0.65	0.31

Table 3: *Averaged F_1 score comparison between using simple unigrams and tf-idfs as they have been described in the text for the prediction of the total CTRS on the* trans *set.*

	uni_T	tf-idf_T	uni_P	tf-idf_P
tot	0.73	**0.86**	0.58	**0.63**

tal score. This behavior can be partially explained by the high correlation between the different codes (Figure 2a).

As a next step, we ran the classifier using the transcribed therapist's text, but having deleted all the instances of those three words. The results, when using the updated tf-idfs and DAs (which yielded the second best overall performance among the interpretable features in the first set of experiments), are presented in Table 4. We can see an overall drop of the F_1 scores corresponding to the tf-idf usage, while, interestingly enough, the performance of the system using the DAs is not affected. When the two feature sets are combined together, the results are even better. This analysis gives cues towards the hypothesis that the DAs could be used to alleviate certain potential problems of an ASR system, which could largely affect the n-grams performance. This issue needs further investigation.

For the *adout* set, we used the pipeline described in Section 3.1. For the role matching module, 3-gram LMs with Witten-Bell smoothing were constructed with IRSTLM [27]. The results, when using the therapist-related tf-idfs, independently or combined with the corresponding DAs, are reported in Table 5. Following the same decision process as with the *trans* set, we selected the $K = 32$ best tf-idfs.

Table 4: *Averaged F_1 score for the binary classification of CBT sessions on the* trans *set using the tf-idfs and the DAs, after having deleted all the instances of the words 'homework', 'agenda', and 'evidence' from the transcripts.*

	tf-idf_T′	da_T′	tf-idf_T′ +da_T′
ag	0.73	0.78	**0.80**
fb	0.69	0.74	**0.78**
un	0.49	0.52	**0.60**
ip	0.46	0.47	0.47
co	0.53	**0.57**	0.56
pt	0.71	0.70	**0.75**
gd	0.66	0.71	**0.74**
cb	0.74	0.75	**0.78**
sc	0.74	0.78	**0.80**
at	0.68	0.75	**0.80**
hw	0.65	0.70	**0.73**
tot	0.71	**0.76**	**0.76**

Table 5: *Averaged F_1 score for the binary classification of CBT sessions on the* adout *set using the therapist's tf-idfs and DAs.*

	tf-idf_T	tf-idf_T +da_T	baseline
ag	0.71	0.71	0.33
fb	0.64	0.62	0.36
un	0.46	0.46	0.46
ip	0.48	0.48	0.48
co	0.45	0.43	0.43
pt	0.60	0.64	0.37
gd	0.63	0.68	0.34
cb	0.67	0.67	0.35
sc	0.61	0.66	0.35
at	0.62	0.64	0.37
hw	0.63	0.65	0.35
tot	0.56	0.58	0.42

An expected overall performance drop is observed, due to potential errors in ASR and to more skewed data (Figure 3). However, tf-idfs yield decent classification results, compared to the baseline performance. The differences when adding the DAs (or other feature sets we tried, without reporting the results here) is statistically insignificant ($p > 0.1$).

6. Conclusions and Future Work

In this work we presented early results for the automatic evaluation of CBT, a widely used psychotherapy approach, showing that even simple linguistic features can lead to good classification performance. Specifically, we demonstrated that the therapist-related features have greater predictive power than the patient-related ones and, among them, the unigrams, under a tf-idf transformation, lead to the best results. However, those are sensitive to very specific words, something that raises questions about finding a set of features more robust to certain ASR errors. Additionally, we showed that those features, despite their good performance for the total CTRS score and most of the individual codes, fail to capture information relevant to the highly imbalanced, human-centric codes, namely the understanding, the interpersonal effectiveness, and the collaboration.

Our future efforts will try to address those issues and will additionally focus on the problem of regression, in order to deduce not only whether a session 'is good' but also 'how good it is'. Finally, we will examine the extent to which different annotation systems (i.e. the ones used in MI and CBT) capture shared vs. unique therapeutic content.

7. Acknowledgements

NF is supported by the USC Annenberg Fellowship. VRM is partially supported by the Mexican Council of Science and Technology (CONACyT).

8. References

[1] E. Proctor, H. Silmere, R. Raghavan, P. Hovmand, G. Aarons, A. Bunger, R. Griffey, and M. Hensley, "Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda," *Administration and Policy in Mental Health and Mental Health Services Research*, vol. 38, no. 2, pp. 65–76, 2011.

[2] D. Bone, C.-C. Lee, T. Chaspari, J. Gibson, and S. Narayanan, "Signal processing and machine learning applications." vol. 34, no. 5, pp. 189–196, 2017.

[3] B. Xiao, D. Can, P. Georgiou, D. Atkins, and S. Narayanan, "Analyzing the language of therapist empathy in motivational interview based psychotherapy," in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2012, pp. 1–4.

[4] D. Atkins, M. Steyvers, Z. Imel, and P. Smyth, "Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification," *Implementation Science*, vol. 9, no. 1, p. 49, 2014.

[5] J. Gibson, N. Malandrakis, F. Romero, D. Atkins, and S. Narayanan, "Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms," in *Proc. Interspeech*, 2015, pp. 1947–1951.

[6] D. Can, D. Atkins, and S. Narayanan, "A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations," in *Proc. Interspeech*, 2015, pp. 339–343.

[7] B. Xiao, Z. Imel, D. Atkins, P. Georgiou, and S. Narayanan, "Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling," in *Proc. Interspeech*, 2015, pp. 2489–2493.

[8] B. Xiao, D. Bone, M. Segbroeck, Z. Imel, D. Atkins, P. Georgiou, and S. Narayanan, "Modeling therapist empathy through prosody in drug addiction counseling," in *Proc. Interspeech*, 2014, pp. 213–217.

[9] S. Lord, E. Sheng, Z. Imel, J. Baer, and D. Atkins, "More than reflections: empathy in motivational interviewing includes language style synchrony between therapist and client," *Behavior therapy*, vol. 46, no. 3, pp. 296–303, 2015.

[10] M. Tanana, K. Hallgren, Z. Imel, D. Atkins, P. Smyth, and V. Srikumar, "Recursive neural networks for coding therapist and patient behavior in motivational interviewing." in *Proc. Computational Linguistics and Clinical Psychology (CLPsych)*, 2015, pp. 71–79.

[11] J. Gibson, D. Can, B. Xiao, Z. Imel, D. Atkins, P. Georgiou, and S. Narayanan, "A deep learning approach to modeling empathy in addiction counseling," in *Proc. Interspeech*, 2016, pp. 1447–1451.

[12] B. Xiao, D. Can, J. Gibson, Z. Imel, D. Atkins, P. Georgiou, and S. Narayanan, "Behavioral coding of therapist language in addiction counseling using recurrent neural networks." in *Proc. Interspeech 2016*, 2016, pp. 908–912.

[13] J. Gibson, D. Can, P. Georgiou, D. Atkins, and S. Narayanan, "Attention networks for modeling behaviors in addiction counseling," *Proc. Interspeech*, pp. 3251–3255, 2017.

[14] J. Beck, *Cognitive behavior therapy: Basics and beyond*. New York, NY, USA: Guilford Press, 2011.

[15] T. Creed, S. Frankel, R. German, K. Green, S. Jager-Hyman, K. Taylor, A. Adler, C. Wolk, S. Stirman, S. Waltman *et al.*, "Implementation of transdiagnostic cognitive therapy in community behavioral health: The beck community initiative." *Journal of consulting and clinical psychology*, vol. 84, no. 12, pp. 1116–1126, 2016.

[16] J. Young and J. Beck, *Cognitive therapy scale: Rating manual*. Center for Cognitive Therapy, University of Pennsylvania, Philadelphia, PA, USA: Unpublished manuscript, 1980.

[17] W. R. Miller, T. B. Moyers, D. Ernst, and P. Amrhein, "Manual for the motivational interviewing skill code (misc)," *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*, 2003.

[18] T. B. Moyers, T. Martin, J. K. Manuel, W. R. Miller, and D. Ernst, "The motivational interviewing treatment integrity (miti) code," *Unpublished manuscript*, 2003.

[19] A. Beck, *Cognitive therapy of depression*. New York, NY, USA: Guilford press, 1979.

[20] T. Vallis, B. Shaw, and K. Dobson, "The cognitive therapy scale: psychometric properties." *Journal of consulting and clinical psychology*, vol. 54, no. 3, pp. 381–385, 1986.

[21] B. Xiao, C. Huang, Z. Imel, D. Atkins, P. Georgiou, and S. Narayanan, "A technology prototype system for rating therapist empathy from audio recordings in addiction counseling," *PeerJ Computer Science*, vol. 2, p. e59, 2016.

[22] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986.

[23] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[24] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[25] J. Pennebaker, R. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," University of Texas at Austin, Austin, TA, USA, Tech. Rep., 2015.

[26] N. Malandrakis and S. Narayanan, "Therapy language analysis using automatically generated psycholinguistic norms." in *Proc. Interspeech*, 2015, pp. 1952–1956.

[27] M. Federico, N. Bertoldi, and M. Cettolo, "IRSTLM: an open source toolkit for handling large scale language models." in *Proc. Interspeech*, 2008, pp. 1618–1621.