Approaching Human Performance in Behavior Estimation in Couples Therapy Using Deep Sentence Embeddings

Shao-Yen Tseng[1], *Brian Baucom*[2], *Panayiotis Georgiou*[1]

[1]University of Southern California, Department of Electrical Engineering, USA
[2]The University of Utah, Department of Psychology, USA

shaoyent@usc.edu, brian.baucom@utah.edu, georgiou@sipi.usc.edu

Abstract

Identifying complex behavior in human interactions for observational studies often involves the tedious process of transcribing and annotating large amounts of data. While there is significant work towards accurate transcription in Automatic Speech Recognition, automatic Natural Language Understanding of high-level human behaviors from the transcribed text is still at an early stage of development. In this paper we present a novel approach for modeling human behavior using sentence embeddings and propose an automatic behavior annotation framework. We explore unsupervised methods of extracting semantic information, using *seq2seq* models, into deep sentence embeddings and demonstrate that these embeddings capture behaviorally meaningful information. Our proposed framework utilizes LSTM Recurrent Neural Networks to estimate behavior trajectories from these sentence embeddings. Finally, we employ fusion to compare our high-resolution behavioral trajectories with the coarse, session-level behavioral ratings of human annotators in Couples Therapy. Our experiments show that behavior annotation using this framework achieves better results than prior methods and approaches or exceeds human performance in terms of annotator agreement.

Index Terms: Behavioral signal processing, natural language processing, sequence-to-sequence learning, recurrent neural network

1. Introduction

Mental health is heavily based on observation of human communication. Subjects and patients are evaluated and supported by psychologists towards improved outcomes based on such observations. For example, in couples therapy, therapists identify problems in the couples' behavior. Based on their assessment they encourage mediatory action to improve the relationships of the couples [1]. Currently, domain experts employ subsequent annotation of behavior by multiple expert annotators towards better understanding of the underlying processes. This is applicable to a broad range of psychotherapy disciplines, such as addiction or suicide counseling [2, 3], and can improve future treatments. Unfortunately, multiple expert annotators for behavior cannot be employed in real therapy due to scalability, cost, and delay issues. Instead, the only observation is the real-time observation of the therapist.

The U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick MD 21702-5014 is the awarding and administering acquisition office. This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs through the Psychological Health and Traumatic Brain Injury Research Program under Award No. W81XWH-15-1-0632. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the Department of Defense.

Annotation of human behaviors is usually performed by trained human coders and is an expensive and time-consuming process. As such it only takes place in limited cases for research purposes. Human annotators first have to be trained in accordance with detailed coding manuals [4, 5] to provide accurate and consistent ratings without the influence of personal bias. Trained annotators may then be evaluated to select those with the highest agreements for the final annotation task. The overall process is lengthy and strainful, but even so, agreement in human annotations can still be quite low [5].

The estimation of human behavior is a complex task as it manifests over multiple modalities. The behavioral dimensions of interest to domain experts are also along dependent dimensions. Behavior Signal Processing (BSP) [6–8] seeks to model human behavior through signal processing and machine learning techniques and offers the ability to automatically estimate human behavioral states. This endeavor could prove to be beneficial for reducing annotation time and cost while enhancing psychotherapy quality through more accurate annotations at finer resolutions. Past work has shown the viability of such methods using audio and/or visual features [9–12]. Great advances have also been made in Natural Language Processing (NLP) methods for behavior modeling [13–15]. Can *et al.* [16] studied the use of linguistic features for detecting counseling behavior in Motivational Interviewing. A dynamic model using language models was then proposed by Chakravarthula *et al.* [17] to model transitions in speaker behavior states in sessions of Couples Therapy. An LSTM-based model was then subsequently proposed in [18] which estimated the trajectories of behavior in those sessions.

Previous NLP methods have mainly utilized N-grams for behavior annotation using count-based or RNN language models [17–19]. While N-gram models are suited for modeling language structure they are unable to capture semantic information of entire speech segments. To conceptualize linguistic information from larger context Tanana *et al.* [20] proposed two methods for deriving sentence features. One was a discrete sentence feature model using N-grams and dependency relations in the parse tree. The other was a *recursive* neural network based on word embeddings in addition to the parse tree. These methods sought to encode contextual meaning of sentences into vector representations based on functional relationships of words from the dependency tree.

Encoding sentences into higher levels of abstraction is a form of representation learning which makes it easier to extract relevant information for the task [21]. In this work we present an unsupervised deep learning method for deriving distributed vector representations of sentences that are well-suited for behavior annotation. We explore different methods of training such vectors and demonstrate the benefits of unsupervised training on closely-matched out-of-domain data. We also propose a

comprehensive framework for modeling human behavior using recurrent neural networks (RNN) and the learned deep sentence embeddings. The RNN framework estimates the trajectories of behavior in conversational interactions using sequences of sentence embeddings. Finally, we evaluate our system on behavior ratings from the Couples Therapy Corpus using a regression model on top of the behavior trajectories.

The rest of the paper is organized as follows: Section 2 briefly introduces deep sentence embeddings. The methodology for our behavior annotation framework using the sentence embeddings is described in Section 3. The corpora and learning methods used in all our experiments are described in Section 4. We compare different methods of training deep sentence embeddings and present our experiment results in Section 5. Finally, we conclude and propose future work in Section 6.

2. Deep Sentence Embeddings

Recurrent neural networks have shown great ability in capturing temporal information in sequences by embedding past history information within hidden states in intermediate layers. Using these hidden states the network then makes the best output decision conditioned on this representation of history [22].

Later, it was shown in [23] that significant improvements in NLP tasks such as machine translation could be obtained by embedding the whole history before generating the output. These networks were referred to as sequence-to-sequence models and incorporate an encoder-decoder architecture that encodes the entire input before generating the output at the decoder stage. The power of sequence-to-sequence models in NLP tasks stems from the fact that the structure of language is non-deterministic and highly dependent upon context [23]. By encoding the entire input as an embedding the network learns how to extract relevant information from the whole input before generating the output. In a way, we can say that the hidden states of the encoder represent the contextual concept that is conveyed by the input sentence. These hidden states are sometimes referred to as deep sentence embeddings and have been shown to be more adept in many NLP tasks than knowledge-based or handmade features [24, 25].

3. Methodology

3.1. Deep Conversational Sentence Embeddings

Deep sentence embeddings represent input sentences at a higher or "deeper" level of abstraction. However, the quality of embeddings depends greatly on the training methodology and learning criteria. In this work our goal is to estimate behavior in human interactions, therefore it follows that the sentence embeddings should represent expressed behaviors and conveyed concepts within the conversations. To this end we employ neural conversation models (CM) [26]. These are encoder-decoder networks that have been trained to give responses from queries and are capable of basic conversations. Embeddings from these networks represent real conversations and encode relevant content of the conversation. While they are not explicitly trained to identiy behavior, given the short-term stationarity of behavior we hypothesize that behavioral information is also represented in these embeddings. Therefore, we extract sentence embeddings from conversational encoder-decoder networks to use as input features for behavior annotation. The encoder-decoder architecture is shown in Figure 1 and is described in further detail in Section 4.

3.2. Behavior Annotation

The method described in the previous section is used to generate deep sentence embeddings for each utterance in our dataset.

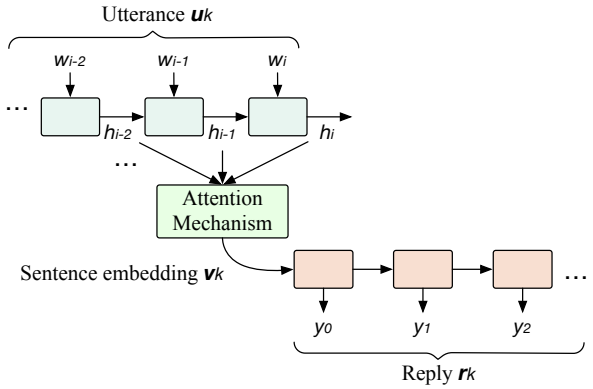

Figure 1: *The encoder-decoder conversation model for generating deep sentence embeddings.*

We then combine multiple utterances into sequences of sentence embeddings. We view these sequences as a representation of conversational information within the interaction over time. Our assumption is that these sentence embeddings generalize information from the speaker in a much richer form than those obtained from the word-level while maintaining temporal information. These sequences of sentence embeddings are therefore ideal for use as features in identifying behavior throughout the interaction.

We apply a sliding window to generate frames of sentence embedding sequences and train an RNN to estimate behavior ratings for each frame. The RNN consists of an LSTM and a feedforward layer in the hidden layers with dropout added after each layer. Since we are estimating a normalized behavior rating between 0 and 1 we use a sigmoid function at the output layer. This architecture is based on that of our previous work [18] which was proven to be effective in modeling behavior with limited training data. Figure 2 shows the architecture of the LSTM-RNN.

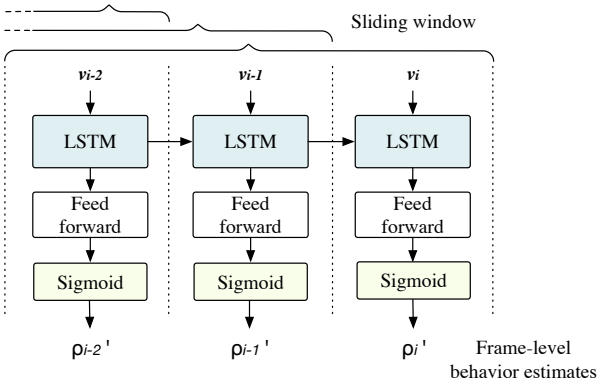

Figure 2: *Architecture of the RNN for estimating frame-level behavior from sequences of sentence embeddings.*

4. Corpora and Learning Methods

4.1. Training Deep Sentence Embeddings

In our experiments we used the OpenSubtitles dataset [27] to train the encoder-decoder model for generating deep sentence embeddings. This dataset contains dialogue from movies which is similar to the back-and-forth structure of the interaction in our target behavior dataset. We processed the data by segmenting paragraphs into sentences. We also removed various generic

expressions such as "*I don't know*" to prevent overtraining on responses that have little relation to queries. Similar to [26] we treat any two consecutive sentences as an utterance-reply pair without considering who uttered the sentence. We then trained the encoder-decoder model to predict the reply given the utterance. However, our work differs in that our final goal is not a working conversation model, but rather a rich semantic representation of the utterance to be used as input feature for behavior annotation. Therefore we want the sentence embeddings to be as compact as possible while still capturing rich contextual information. In our experiments we tried embedding sizes of 100, 500, and 1024 with 3 LSTM layers in the encoder and decoders. We also add an attention mechanism [28] after the encoder to allow the network to focus on more salient portions of the input. The final training set consists of 35 million utterance-reply pairs.

4.2. Behavior Annotation in Couples Therapy

4.2.1. Couples Therapy Corpus

To train behavior models we use data from the UCLA/UW Couple Therapy Research Project [1] which contains audio and video recordings of 134 couples with real marital issues interacting over multiple sessions. In each session, couples discuss a specific topic chosen in turn for around 10 minutes. The behaviors of each speaker are then rated by multiple annotators based on the Couples Interaction [4] and Social Support [29] Rating Systems. The rating system contains 33 behavioral codes such as "Acceptance", "Blame" and "Negativity". Each annotator provides ratings for these codes on a Likert scale of 1 to 9 for every session, where 1 indicates strong absence and 9 indicates strong presence of the behavior. There are 2 to 12 annotators per session with the majority of sessions ($\sim 90\%$) having 3 to 4 annotators. Finally, these ratings are averaged to obtain a 33 dimensional vector of behavior ratings per interlocutor for every session.

In this paper, we focus primarily on the behavior code "Negativity". For our experiments we use manual transcriptions of sessions with mean annotator ratings in the top and bottom 20% of the code range for sessions with good audio quality.

4.2.2. Frame-Level Behavior Metrics

Behavior ratings in the Couples Therapy Corpus are annotated for entire sessions and no labels are provided for individual utterances. However, it is infeasible to treat entire sessions as a single sequence of embeddings due to the high complexity of such a model and data scarcity issues. We also want fine annotations of utterances in addition to session-level annotations. Therefore we employ weakly supervised learning and assign session-level ratings as target values for all short embedding sequences in a session. This assumes that all utterances in a session relate to the overall rating in a non-linear and complex manner and by only considering shorter sequences we can still map back to the session rating.

In our experiments we generate frames of embedding sequences using a sliding window of 3 utterances with a shift of 1 utterance. The RNN takes these embedding sequences as input and is trained to predict the session-level rating using an SGD optimizer. The result is an overlapping trajectory of frame-level behavior metrics over the session. We refer to these values as metrics since they convey information of behavior and indirectly relate to the session rating in some form.

4.2.3. Annotating Behavior

Since we do not have annotations for individual utterances to compare with, we validate our system with session-level ratings. We do this by fusing, using the techniques described below, the frame-level behavior metrics to derive an estimate for the session-level score. In a sense this method is similar to the human process of integrating behavioral information over time to arrive at a gestalt opinion of the session.

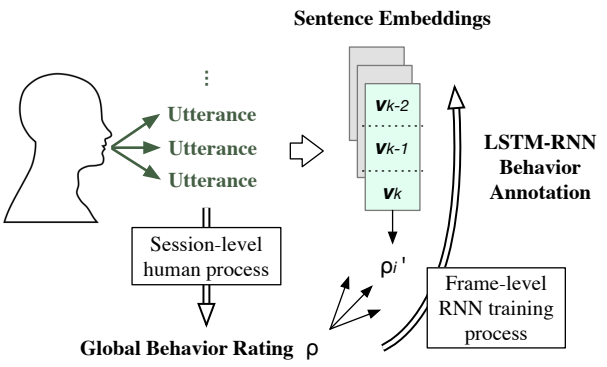

Figure 3: *The behavior annotation framework. Frame-level annotations are trained with global ratings using weakly supervised learning.*

To compare results we apply the fusion method used in our previous work [18]. Specifically, we used an RBF-Kernel Support Vector Regressor to learn a mapping from the median of the frame-level behavior metrics in a session to the true rating. At test time, we apply this map on the median of the behavior metrics to obtain an estimated rating for the entire session. Although there are many different fusion techniques we implement this method for consistency with prior work. Session-level fusion is not the focus of this study. An overview of our proposed behavior annotation framework is shown in Figure 3.

5. Experimental Evaluation

In all our experiments we used a leave-one-couple-out cross-validation scheme to separate the Couples Therapy data into train and test sets. For each fold data from one couple is held out from training and used for evaluation. This resulted in a total of 134 folds. The approach of our experiments are as follows:

- Train a conversation model using an encoder-decoder architecture by predicting replies to utterances in the OpenSubtitles dataset. This first step is domain-data independent and is only done once. The following steps are run on the per-fold split.
- Extract deep sentence embeddings for all utterances in the Couples Therapy Corpus. Use the attention layer of the conversation model as an embedding.
- Use a sequence of sentence embeddings as features and train an RNN to estimate the session-level ratings from each embedding sequence. This is the first supervised step.
- To obtain session-level behavior ratings we train an RBF-Kernel SVR to map the median of frame-level behavior metrics to a final score for each fold.

Table 1: *MAE of estimated ratings using different models.*

Model	MAE
Word-level sequences [18]	1.53
Sum W2V	1.43
NMT embeddings	1.68
CM embeddings	**1.37**

Table 2: *Comparison of inter-annotator agreement using Krippendorff's alpha.*

Annotator Configuration	Krippendorff's Alpha			
All human annotators	0.821			
Random replacement with random predictions	0.492			
	Word sequences [18]	Sum W2V	NMT embeddings	CM embeddings
Random replacement with machine predictions	0.7739	0.7776	0.7511	0.7832
Outlier replaced with machine prediction	0.8249	0.8368	0.8010	**0.8403**

5.1. Predicting True Behavior Ratings

For validation we compared the estimated behavior rating from the fusion outputs to scores given by human annotators. We trained different RNNs for behavior annotation with various types of embedding sequences for comparison. These included our previous work of word-level sequences [18], the sum of *word2vec* [30] embeddings in a sentence, and deep sentence embeddings extracted from an English-to-French neural machine translation (NMT) model [23]. For fair comparison the dimensions of the embeddings were fixed to the same size.

It is important to note that there is no absolute truth in the reference annotations. These are subjective, and human annotators have disagreements. As such we can at best achieve to reach agreement with the mean comparable to the inter-annotator agreement. We thus have two validation metrics: *(1)* The Mean Absolute Error (MAE) with the average rating, that we know is not necessarily the gold standard; and *(2)* Treat our system as another annotator and see how it compares to existing human expert annotators in terms of inter-annotator agreement.

Table 1 shows the Mean Absolute Error (MAE) between estimated ratings of different models and the average score of human annotators. Our proposed model using deep sentence embeddings performs significantly better than prior work [18] (*Mann-Whitney U-test,* $p < 0.05$).

To evaluate inter-annotator agreement we mixed our estimated ratings with human annotations and calculated Krippendorff's alpha coefficient for two different configurations. In the first configuration we randomly replaced one human annotation with estimated ratings. We found that while all models achieved lower inter-annotator agreement than human-only annotations, the system trained on embedding sequences from the conversation model gave the best results. Next, we selected human annotations that deviated most from the mean and replaced them with estimated ratings. Again we found that our proposed method gave the highest agreement and even outperformed outlier human annotators in terms of agreement with other annotators. Table 2 shows the inter-annotator agreement between estimated ratings and human annotators under the different configurations. The Krippendorff's alpha for random replacement with random values is also shown as reference.

5.2. Rating Behavior in Text

To see how our system performs on out-of-domain data we rated negativity on dialogue from the television series "Friends" as an example. We used transcripts from the show and tokenized each speaker turn into one or more utterances. To track behavior in the overall interaction we assumed that all utterances originated from a single person and applied the sliding window on all the data. Even though "Friends" is a comedy and is expected to be mostly positive, our system was able to identify many utterances that seem to exhibit negative behavior. Some examples of negativity in the dialogues are shown in Table 3. These results are encouraging in that they show how our behavior annotation framework is able to learn from weak labels and be transferable to other domains.

Table 3: *Examples of negativity in utterances from Friends.*

Less Negative Sentences
Alright, this barbecue is gonna be very fun
I'm not saying he has to spend the whole evening with me, but at least check in .
More Negative Sentences
I'm the girl in the veil who stomped on your heart in front of your entire family.
Joey, this is sick, it's disgusting, it's not really true, is it?

6. Conclusions and Future Work

In this work we proposed a behavior annotation framework based on deep sentence embeddings trained using neural conversation models. We theorize that sentence embeddings from conversation models are more adept at capturing conversational concepts which relate better to behavior. We then modeled interactions using sequences of these embeddings and trained an LSTM-RNN to estimate trajectories of behavior in Couples Therapy Sessions. Finally, we evaluated our system by fusing local behavior metrics into a session-level rating and compared with human annotations. The results of our experiments showed that using embedding sequences from conversation models as input features for behavior modeling achieves higher inter-annotator agreement with human annotators over other types of sentence embeddings. Such an approach gives session-level behavior ratings close to human annotators and even outperforms outlier humans.

Our system seeks to alleviate the expensive and time-consuming process of manual behavior annotation required for observational studies in psychotherapy. In addition, through weakly supervised learning, we provide objective behavior ratings at a finer resolution of per utterance. The utterance-level behavior ratings are more capable than previous works at capturing behavior trajectories in a couples interaction session and allow for more detailed analysis by psychologists.

Future work includes the study of dyadic behavior models which considers the interplay of behavioral states of both speakers. We also plan on incorporating information from other modalities such as acoustic or video for a more complete and detailed model. Finally, we hope to employ our model in assisting with behavioral evaluations in a wide range of domains and applications, such as in suicide prevention.

7. References

[1] A. Christensen, D. Atkins, S. Berns, J. Wheeler, D. Baucom, and L. Simpson, "Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples," *Journal of Consulting and Clinical Psychology*, vol. 72, no. 2, pp. 176–191, 2004.

[2] D. Can, D. Atkins, and S. S. Narayanan, "A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations," in *Proceedings of Interspeech*, Sep. 2015.

[3] C. J. Bryan, M. David Rudd, E. Wertenberger, N. Etienne, B. N. Ray-Sannerud, C. E. Morrow, A. L. Peterson, and S. Young-McCaughon, "Improving the detection and prediction of suicidal behavior among military personnel by measuring suicidal beliefs: an evaluation of the Suicide Cognitions Scale," *J Affect Disord*, vol. 159, pp. 15–22, Apr 2014.

[4] C. Heavey, D. Gill, and A. Christensen, "Couples interaction rating system 2 (cirs2)," *University of California, Los Angeles*, vol. 7, 2002.

[5] G. Margolin, P. H. Oliver, E. B. Gordis, H. G. O'hearn, A. M. Medina, C. M. Ghosh, and L. Morland, "The nuts and bolts of behavioral observation of marital and family interaction," *Clinical child and family psychology review*, vol. 1, no. 4, pp. 195–213, 12 1998.

[6] M. Black, A. Katsamanis, C.-C. Lee, A. Lammert, B. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Automatic classification of married couples' behavior using audio features," in *In Proceedings of InterSpeech*, Makuhari, Japan, Sep. 2010.

[7] P. G. Georgiou, M. P. Black, and S. S. Narayanan, "Behavioral signal processing for understanding (distressed) dyadic interactions: some recent developments," in *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*. Scottsdale, AZ: ACM, 2011, pp. 7–12.

[8] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. PP, no. 99, pp. 1 –31, 2013.

[9] J. Gibson, B. Xiao, P. G. Georgiou, and S. S. Narayanan, "An audio-visual approach to learning salient behaviors in couples' problem solving discussions," in *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME)*, Jul. 2013.

[10] B. Xiao, P. Georgiou, B. Baucom, and S. Narayanan, "Power-spectral analysis of head motion signal for behavioral modeling in human interaction," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, May 2014.

[11] H. Li, B. Baucom, and P. Georgiou, "Sparsely connected and disjointly trained deep neural networks for low resource behavioral annotation: Acoustic classification in couples' therapy," in *Proceedings of Interspeech*, San Francisco, CA, September 2016.

[12] 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 *Proceedings of Interspeech*, Dresden, Germany, September 2015.

[13] P. G. Georgiou, M. P. Black, A. Lammert, B. Baucom, and S. S. Narayanan, ""That's aggravating, very aggravating": Is it possible to classify behaviors in couple interactions using automatically derived lexical features?" in *Proceedings of Affective Computing and Intelligent Interaction (ACII), Lecture Notes in Computer Science*, Oct. 2011.

[14] B. Xiao, P. Georgiou, Z. E. Imel, D. Atkins, and S. Narayanan, ""Rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing," *PLOS ONE*, December 2015.

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

[16] D. Can, Z. Imel, D. Atkins, P. Georgiou, and S. Narayanan, "'it sounds like...': A natural language processing approach to detecting counselor reflections in motivational interviewing," *Journal of Counseling Psychology*, 2015.

[17] S. N. Chakravarthula, R. Gupta, B. Baucom, and P. Georgiou, "A language-based generative model framework for behavioral analysis of couples' therapy," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, Apr. 2015.

[18] S.-Y. Tseng, S. N. Chakravarthula, B. Baucom, and P. Georgiou, "Couples behavior modeling and annotation using low-resource LSTM language models," in *Proceedings of Interspeech*, San Francisco, CA, September 2016.

[19] 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 *Proceedings of Interspeech*, San Francisco, CA, September 2016.

[20] M. Tanana, K. A. Hallgren, Z. E. Imel, D. C. Atkins, and V. Srikumar, "A comparison of natural language processing methods for automated coding of motivational interviewing," *Journal of substance abuse treatment*, vol. 65, pp. 43–50, 2016.

[21] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[22] M. Boden, "A guide to recurrent neural networks and backpropagation," *the Dallas project*, 2002.

[23] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[24] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents." in *ICML*, vol. 14, 2014, pp. 1188–1196.

[25] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, and R. Ward, "Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 4, pp. 694–707, 2016.

[26] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[27] J. Tiedemann, "News from opus-a collection of multilingual parallel corpora with tools and interfaces," in *Recent advances in natural language processing*, vol. 5, 2009, pp. 237–248.

[28] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[29] J. Jones and A. Christensen, "Couples interaction study: Social support interaction rating system," University of California, Los Angeles, Technical manual, 1998.

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *In Proceedings of Workshop at ICLR*, 2013.