Behavioral Coding of Therapist Language in Addiction Counseling using Recurrent Neural Networks

Bo Xiao[1], *Doğan Can*[1], *James Gibson*[1], *Zac E. Imel*[2],
David C. Atkins[3], *Panayiotis Georgiou*[1], *Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA, USA
[2]Department of Educational Psychology, University of Utah, Salt Lake City, UT, USA
[3]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA
[1]sail.usc.edu, [2]zac.imel@utah.edu, [3]datkins@u.washington.edu

Abstract

Manual annotation of human behaviors with domain specific codes is a primary method of research and treatment fidelity evaluation in psychotherapy. However, manual annotation has a prohibitively high cost and does not scale to coding large amounts of psychotherapy session data. In this paper, we present a case study of modeling therapist language in addiction counseling, and propose an automatic coding approach. The task objective is to code therapist utterances with domain specific codes. We employ Recurrent Neural Networks (RNNs) to predict these behavioral codes based on session transcripts. Experiments show that RNNs outperform the baseline method using Maximum Entropy models. The model with bi-directional Gated Recurrent Units and domain specific word embeddings achieved the highest overall accuracy. We also briefly discuss about client code prediction and comparison to previous work.

Index Terms: Recurrent neural network, behavioral coding, motivational interviewing, language modeling

1. Introduction

Promoting mental healthcare is an important societal need, particularly for the prevention of drug and alcohol abuse [1]. Psychotherapy research investigates factors that contribute to the effectiveness of treatment as well as methods of assessing therapist competence. Observational coding of therapist and client behaviors using domain knowledge inspired coding manuals has been a standard approach in research [2]. Similar to dialog act tagging, psychotherapy coding often focuses on discrete linguistic behaviors and classifies them in categories of clinical interest.

In this paper, we take *Motivational Interviewing* (MI) as an example, which is a type of psychotherapy treatment [3]. MI is clinically effective and widely adopted in applications such as addiction counseling. It emphasizes the intrinsic motivation of clients to change their addictive behavior [4]. The *Motivational Interviewing Skill Code* (MISC) — a widely used coding manual — classifies each utterance into one of a set of mutually exclusive and exhaustive codes, such as *Facilitate*, *Simple Reflection*, *Open Question*, *etc.* [5]. The MISC coding system attempts to capture the task relevant aspects of utterances. Statistics of the codes assigned to an MI session are used as measures of the therapist's treatment fidelity. Here one session corresponds to one appointment of MI treatment or intervention, conducted in the form of therapist-client conversation.

Traditionally, human coders manually observe session recordings and make code assignments. However, the coding process is costly in both time and human resources [6]. Coder training and reliability evaluation are additional critical issues besides the cost of coding. These limitations make large scale applications of manual coding unrealistic [7]. To address this difficulty, computational methods of analyzing human behaviors have been proposed to complement human experts' judgments. Multimodal signal processing and machine learning methods have shown promise in modeling behavioral cues and their links to expert judgments [8, 9].

Can *et al.* proposed the first computational model towards identifying *Reflections*, a major class of codes in MISC [10]. Maximum entropy Markov model with word N-gram features and code context achieved the best performance. Atkins *et al.* employed labeled topic models to predict a set of 12 MISC codes on talk turn and psychotherapy session levels [11]. They compared coder-coder *vs.* computer-coder agreement on each code. Can *et al.* proposed a method using conditional random field to model the sequence of MISC codes, which predicted expert judgments on the full MISC code set [12]. They incorporated word N-grams, MISC codes, speaker roles (therapist and client) and context from neighboring utterances in the feature functions. Tanana *et al.* proposed two competing methods to predict MISC codes for each utterance [13, 14]. One of the proposed methods was a large multinomial regression model based on word N-grams and dependency relations in the parsing tree of an utterance. The other was a recursive neural network model (to be distinguished from recurrent neural network) based on the parsing tree and word embeddings. They found that the two methods were comparable in performance while the regression model was slightly better. In general, previous work has shown that prediction accuracy for a code relates to the degree of the code's sparsity and the level of human agreement [11].

In this work, our main focus is on therapist code prediction, although we also briefly discuss predicting client codes. For simplicity, we treat each utterance as an independent sample without context. We first set up the Maximum Entropy (MaxEnt) model as a baseline for MISC code prediction. The MaxEnt model is in the family of log-linear models, which have achieved good performance in a wide range of natural language processing tasks [15, 16]. We also choose this model because similar methods have been used in previous work on MISC code prediction [10, 12].

We then employ Recurrent Neural Networks (RNNs) in a deep learning framework and examine their performances in comparison to the baseline. Specifically, we use Long Short-Term Memory (LSTM) [17] and Gated Recurrent Unit (GRU)

[18] RNNs in order to address the "vanishing gradient problem" associated with simple RNNs. These models have achieved state-of-the-art results in a number of machine learning tasks [19, 20]. We design the network architecture as either consuming the input text stream only in the forward direction or in both the forward and backward directions. We represent words with word embeddings, which are generated by two types of "word-to-vector" transformations — either trained on domain relevant data or generic data [21]. We compare the results achieved by different combinations of the above network units, architectures, and word embeddings.

2. Method

We assume each utterance is represented by a word sequence $\mathbf{w} = \{w_0, w_1, \cdots, w_{L-1}\}$, where L is the number of words in the utterance. We then assume a function $c = f(\mathbf{w})$ maps $\mathbf{w}$ to a MISC code $c \in \{1, 2, \cdots, C\}$, with C being the count of defined code types. Our goal is to find the function f^* minimizing the error between the predicted and expert annotated codes.

2.1. Maximum Entropy Model

The MaxEnt model [16] derives the posterior probability $P(c|\mathbf{w})$ based on a group of feature functions $f_i(\mathbf{w}, c)$, as shown in (1). Here λ and $Z(\mathbf{w})$ denote the weights and the partition function, respectively. The predicted code c^* is the one maximizing $P(c|\mathbf{w})$.

$$P(c|\mathbf{w}) = \frac{1}{Z(\mathbf{w})} exp\left(\sum_i \lambda_i f_i(\mathbf{w}, c)\right) \quad (1)$$

In this work, the feature functions are simple word N-gram (up to tri-gram) counts. An example feature function is shown in (2).

$$f_i(\mathbf{w}, c) = \sum_{j=0}^{L-2} g_i((w_j, w_{j+1}), c) \quad (2)$$

The function g_i describes the case that a bi-gram pattern $(\widehat{p}, \widehat{q})$ is associated with code $\widehat{c}$. $g((p, q), c) = 1$ if $(p, q) = (\widehat{p}, \widehat{q})$ and $c = \widehat{c}$, otherwise it equals 0. The feature function $f_i(\mathbf{w}, c)$ then counts the appearances of $(\widehat{p}, \widehat{q})$ in $\mathbf{w}$ if $c = \widehat{c}$. The weights λ are learned on the training set. We use the L-BFGS algorithm [22] for optimization and the MaxEnt toolkit in [23] for implementation.

2.2. Recurrent Neural Networks

We employ two types of neural network architectures — either consuming inputs only in the forward direction (uni-directional) or in both the forward and backward directions (bi-directional), as demonstrated in Fig. 1 and Fig. 2, respectively.

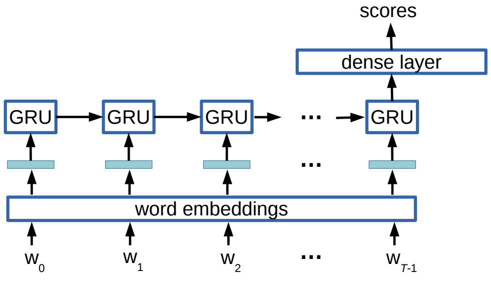

Figure 1: *Uni-directional neural network architecture.*

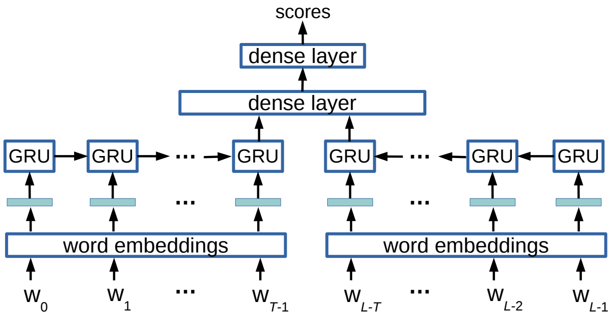

Figure 2: *Bi-directional neural network architecture.*

The bottom layer of the network is a word-embedding layer. Word embeddings represent words in a lower dimensional, continuous vector space, instead of integer indices or vectors in the dimensionality of the vocabulary [21]. The word vector space is not only an efficient data representation, but also captures semantic relations between words such that words close in meaning are also close in the space. The word-embedding layer is initialized using a pre-trained model, which was trained on a large text corpus in an unsupervised manner based on word co-occurrence statistics. This layer is then fine-tuned during training. We examine two sources for pre-training the word embeddings, either from in-domain or generic text data. Specifically, we use the gensim [24] toolkit for training in-domain word embeddings, and use the word embeddings generated on Wikipedia and Gigaword-5 data using the GloVe toolkit [25] as generic word embeddings.

The recurrent layer takes one word vector at a time as its input, and passes down its internal states. We take the output of the recurrent unit at the last time instance as the layer's output, because we are only interested in the complete meaning of an utterance. We examine two types of recurrent units, namely LSTM [17] and GRU [18]. Both of them have gated structures to control input and output flows, in order to eliminate the "vanishing gradient" in error back-propagation during training. This enables a network to capture long-distance relations in sequential data. In this work, we threshold word sequence lengths (denoted as T in Fig. 1 and Fig. 2) to 50, which covers more than 99% of utterances in the dataset. Longer utterances are truncated, while shorter ones are padded with zeros. In the bi-directional network, the utterance is first reversed then truncated for the backward flow. We set a dropout rate of 0.2 for the recurrent layer to prevent overfitting.

At the top of the uni-directional network, a fully connected layer of hidden neurons (*i.e.,* dense layer) takes the output vector of the recurrent layer, and maps it to a vector of scores in the dimension of MISC codes. In the training phase, the score vector is set to 1 in the corresponding dimension of the MISC code, and 0 elsewhere. We use softmax activation for the output of the dense layer and train the network to optimize the categorical cross-entropy loss function using ADAM algorithm [26]. In the testing phase, the code associated with the highest score is selected as the prediction. For the bi-directional network, we add another dense layer between the recurrent layer and the top layer. This layer concatenates the outputs from the forward and backward recurrent units, and allows them to interact before the top layer. We also set a dropout rate of 0.2 for this layer. We use the Keras toolkit [27] with the Theano [28] back-end for implementation.

3. Data Corpus

The domain relevant data come from six large scale MI studies. Five of them are intervention studies targeting alcohol abuse by young people (ARC, ESPSB, ESP21), marijuana abuse (iCHAMP), and poly-drug abuse (HMCBI) [11]. Due to resource constrains, 148 out of a total of 899 sessions were randomly selected for manual coding. The sixth study is aimed at therapist training and involves real and standardized patients (*i.e.,* role-played) [29]. Of the 826 sessions, 195 were randomly selected for manual coding. Sessions last from 20 minutes to 1 hour.

The selected sessions were manually transcribed with talk-turn level time stamps. The transcripts include verbal and non-verbal vocal behaviors such as disfluencies, laughters, back-channels, and overlapped speech. Trained human coders conducted MISC coding based on the audio recordings and transcripts. Each turn may contain multiple codes; coders make subjective judgments on the necessity to segment a turn into multiple utterances, each having a complete thought and a unique code. Some sessions were MISC coded more than once to measure inter- and intra-coder reliability, while the majority were coded only once. The ratio of utterances having coder disagreement is less than 4%; we thus randomly pick one of the assigned codes for each utterance as the reference label.

Table 1: *MISC code grouping and counts in the dataset.*

Code	Original MISC code	Count
Therapist		
FA	Facilitate	15973
GI	Giving information	18120
RES	Simple reflection	6390
REC	Complex reflection	4053
QUC	Closed question (Yes/No)	6343
QUO	Open question (Wh- type)	5597
MIA	MI adherent: Affirm; Reframe; Emphasize control; Support; Filler; Advice with permission; Structure; Raise concern with permission	5984
MIN	MI non-adherent: Confront; Direct; Advice without permission; Warn; Raise concern without permission	1299
Client		
FN	Follow/Neutral	52333
POS	Change talk: positive valence of Reason; Commitment; Taking steps; Other	6630
NEG	Sustain talk: negative valence of the previous row	6218

The original MISC code set contains 28 codes. Some of them are too sparse in the data to support statistically meaningful models. Can *et al.* proposed grouping 19 therapist codes into 7 categories, and all 9 client codes into a single category to address this issue [12]. Six of the seven therapist categories are identical to the original MISC codes, while the last one covers all the remaining codes. In this work, we split the last category according to whether the code represents MI adherent or non-adherent behavior, *i.e.,* whether it follows the spirit of MI. We categorize client codes according to the valence of changing, sustaining, or being neutral to the addictive behavior [11]. Therefore, we have 8 and 3 classes for therapist and client code prediction, respectively. For the experiments, we employ a subset of utterances that do not have overlapped speech, taking about 75% of all utterances. The grouped and original MISC codes along with the grouped counts in the dataset for the experiments are summarized in Table 1.

We split the data into training/testing parts by sessions with roughly 2:1 ratio. The split is speaker independent, *i.e.,* therapists and standardized patients in training do not appear in test. For tokenization, we remove punctuations except apostrophes, replace underscores with spaces, normalize non-verbal vocalizations into “laughter” or “vocal noise”, and finally lowercase all text. An additional text corpus of psychotherapy transcripts (called “general psychotherapy corpus”, 6.5M words) is added to the word-embedding training [30]. Data sizes for therapist/client are summarized in Table 2.

Table 2: *Session, utterance, and word counts in train/test splits.*

Subject	Sessions	Utterances	Words
Therapist	236 / 101	41236 / 22523	444K / 237K
Client	236 / 101	42330 / 22851	475K / 258K

4. Experimental Results

4.1. Therapist Code Prediction

For the MaxEnt approach, we found the combination of uni-, bi- and tri-gram features yielded the highest accuracy. For the RNN approach, we employed the general psychotherapy corpus and the MISC utterances, except the therapist samples from the test set, to train the in-domain word embeddings. In RNN training, the last 10% of the training utterances were used as a validation set. Word vector dimensionality of 100 and 200 were the best for in-domain and generic word embeddings, respectively. For the in-domain word-embedding, CBOW method was superior to skip-gram. The recurrent units had internal dimensionality of 256. The dense layers on top of the uni- and bi-directional networks had 256 input dimensions. The dense layer in the middle of the bi-directional network had 512 and 256 input and output dimensions, respectively. We trained the RNNs with an early-stop strategy, *i.e.,* if the current epoch does not reduce the loss function or prediction error on the validation set, then the previous epoch is considered final.

The MaxEnt model obtained an overall accuracy (*i.e.,* percentage of correct classifications) of 72.17% on therapist code prediction. In Table 3 we report the overall accuracies by the RNNs. We can see that the RNNs exceeded the MaxEnt model. The best result of 75.03% was achieved using in-domain word embeddings and GRU units in a bi-directional network, which was a 2.86% absolute improvement over the MaxEnt baseline. Kappa values of agreement to the target labels for the best MaxEnt and RNN results are 0.652 and 0.686, respectively. In general, GRU, in-domain word embeddings, bi-directional architecture outperformed LSTM, generic word embeddings, uni-directional architecture, respectively. GRU was more computationally efficient compared to LSTM — the training epochs for the latter were about 1.8 times slower. Training in most configurations finished in less than 10 epochs. Bi-directional setups generally required fewer epochs, though each epoch was about 2 times slower.

Table 4 shows the confusion matrix of target codes and the best RNN predictions. FA is well separated from others,

Table 3: *Overall therapist code prediction accuracies (percentage) by the RNNs.*

	Uni-directional		Bi-directional	
Word-embedding	LSTM	GRU	LSTM	GRU
In-domain	74.67	74.75	74.75	**75.03**
Generic	73.48	73.91	73.55	73.55

The best accuracy by the MaxEnt model is 72.17%.

which covers mostly phrases to keep the conversation going, such as “okay” and “yeah”. RES and REC are confused with each other due to their subtle difference, *i.e.,* whether the therapist adds meaning or emphasis when reflecting the client’s statement. QUC and QUO are relatively well predicted possibly due to their linguistic structures. QUC is confused with RES possibly because RES could be carried out in a question form. There are some confusions between QUC and QUO. Though typically QUC and QUO are “Yes/No” and “Wh- type” questions, coding is based on the semantic meaning concerning whether it looks for a specific answer or invokes the client to tell more (*e.g.,* “where do you live?” warrants a QUC code). MIA and MIN are relatively well separated from each other, which is desirable for treatment fidelity assessment.

GI, the largest class, is confused with all the others. GI and reflections are mistaken for each other possibly due to the lack of context, *e.g.,* the previous client statement. GI and questions (especially QUC) are confusing likely due to the lack of prosodic information, *e.g.,* a rising pitch in the end may indicate a question instead of a statement. GI and MIA, MIN are confused likely due to the subtleness of attitude — GI marks a neutral valence of providing information and educating the client; MIA indicates a motivating and supportive attitude; and MIN indicates a directive and critical attitude. Prosodic information may potentially help distinguishing GI and MIA, MIN. Errors on MIN are also partly due to sparse training samples in this class.

Table 4: *Confusion matrix of therapist code prediction using RNN (in-domain word embeddings, bi-directional GRUs). Rows represent manual coding; columns represent predictions.*

	FA	GI	RES	REC	QUC	QUO	MIA	MIN
FA	5606	47	22	0	20	4	56	0
GI	149	5529	419	135	152	55	325	27
RES	34	578	1015	230	118	24	84	4
REC	3	350	409	514	49	12	40	4
QUC	23	202	202	21	1494	199	27	4
QUO	6	79	23	5	158	1666	20	1
MIA	101	714	104	62	35	27	1062	3
MIN	4	173	17	18	26	6	15	12

4.2. Client Code Prediction

The best performing setup of MaxEnt and RNN models achieved overall accuracies of 80.77% and 82.39% for predicting client codes, respectively (word embeddings pre-trained on all but client test utterances plus the general psychotherapy corpus). However, as shown in Table 1 client codes were highly biased to FN. Neither of the two models exceeded the chance level of 82.78%.

Client change/sustain-talk are arguably more important behaviors. To better detect them, errors on these codes may have higher weights in the loss function in RNN training. We set the weight to 1.0 for FN, while sample the weights from 1.0 to 4.0 for POS/NEG. In Table 5 we report the F1 scores and Kappa values. As the weight becomes higher, F1 scores of POS/NEG and the Kappa value increase while the F1 score for FN decreases. However, the problem is still challenging. Disproportion of codes and the lack of contextual modeling may be reasons for the gap.

Table 5: *F1 scores and Kappa in client code prediction.*

Method	F1 score			Kappa
	FN	POS	NEG	
MaxEnt	0.897	0.234	0.264	0.224
RNN, 1.0	0.906	0.174	0.218	0.180
RNN, 2.0	0.900	0.245	0.214	0.214
RNN, 3.0	0.888	0.276	0.258	0.253
RNN, 4.0	0.870	0.300	0.286	0.272

4.3. Empirical Comparison With Previous Work

We review performances in the previous work (mentioned in Section 1), and empirically compare them with the current results, shown in Table 6. Note that because of differences on code/data selection, exact comparisons are not available. RNN and CRF are comparable though context information is not used in the current RNN model. Gaps between human-RNN and human-human agreements are larger for POS, NEG, and REC.

Table 6: *Comparison of code prediction performances.*

	F1 score			Kappa	
Code	DSF [13]	CRF [12]	RNN	Human [11]	RNN
FA	0.94	0.94	0.96	-	0.95
GI	0.69	0.74	0.76	0.76	0.65
RES	0.48	0.49	0.47	0.52	0.42
REC	0.39	0.45	0.43	0.61	0.40
QUC	0.68	0.72	0.71	0.76	0.68
QUO	0.77	0.81	0.84	0.86	0.83
POS	0.29	-	0.30	0.63	0.21
NEG	0.27	-	0.29	0.66	0.22

5. Conclusion

Automatic coding of therapist and patient behaviors is pivotal to scaling up psychotherapy research and treatment fidelity assessment. Computational methods have been proposed to predict expert judgments on MISC codes given the word sequences. In this paper, we have constructed RNN models to predict expert-annotated MISC codes. Experimental results demonstrate that RNNs achieve better performance than MaxEnt models for predicting therapist codes. In particular, the network with in-domain word embeddings and bi-directional GRUs offered the best performance. Improvement on client code prediction is obtained but the problem is still challenging.

In the future, we plan to incorporate prosodic features to better discriminate codes having similar lexical forms. The alignment between words and the speech signal can be derived by force-alignment using an automatic speech recognizer. Context modeling and hierarchical neural network architectures may be helpful to capture dependencies on the utterance and turn levels. The long term goal is to fully automate the system by using ASR output for code prediction [31].

6. References

[1] Substance Abuse and Mental Health Services Administration, *Results from the 2012 National Survey on Drug Use and Health: Summary of National Findings. NSDUH Series H-46, HHS Publication No. (SMA) 13-4795. Rockville, MD, U.S.A.*, 2013.

[2] C. S. Schwalbe, H. Y. Oh, and A. Zweben, "Sustaining motivational interviewing: a meta-analysis of training studies," *Addiction*, vol. 109, no. 8, pp. 1287–1294, 2014.

[3] W. R. Miller and S. Rollnick, *Motivational interviewing: Helping people change*. Guilford Press, 2012.

[4] W. R. Miller and G. S. Rose, "Toward a theory of motivational interviewing," *American Psychologist*, vol. 64, no. 6, p. 527, 2009.

[5] 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.

[6] T. B. Moyers, T. Martin, J. K. Manuel, S. M. Hendrickson, and W. R. Miller, "Assessing competence in the use of motivational interviewing," *Journal of substance abuse treatment*, vol. 28, no. 1, pp. 19–26, 2005.

[7] Z. E. Imel, M. Steyvers, and D. C. Atkins, "Computational psychotherapy research: Scaling up the evaluation of patient-provider interactions," *Psychotherapy*, vol. 52, no. 1, p. 19, 2015.

[8] S. Narayanan and P. Georgiou, "Behavioral Signal Processing: Deriving Human Behavioral Informatics from Speech and Language," *Proceeding of IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[9] B. Xiao, Z. E. Imel, P. Georgiou, D. C. Atkins, and S. S. Narayanan, ""rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing," *PloS one*, vol. 10, no. 12, p. e0143055, 2015.

[10] D. Can, P. Georgiou, D. Atkins, and S. S. Narayanan, "A Case Study: Detecting Counselor Reflections in Psychotherapy for Addictions using Linguistic Features," in *Proc. Interspeech*, Portland, Sep. 2012, pp. 2254–2257.

[11] D. C. Atkins, M. Steyvers, Z. E. 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.

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

[13] 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*, 2016, in press.

[14] 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," *NAACL HLT 2015*, p. 71, 2015.

[15] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Computational linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[16] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modelling," *Computer Speech & Language*, vol. 10, no. 3, pp. 187–228, 1996.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint*, p. arXiv:1412.3555, 2014.

[19] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. Interspeech*, 2014, pp. 338–342.

[20] S. Ravuri and A. Stolcke, "Recurrent neural network and lstm models for lexical utterance classification," in *Proc. Interspeech*, 2015, pp. 135–139.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, http://arxiv.org/abs/1301.3781.

[22] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, no. 1-3, pp. 503–528, 1989.

[23] L. Zhang, *Maximum Entropy Modeling Toolkit for Python and C++*, 2013, https://github.com/lzhang10/maxent.

[24] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proc. LREC Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[25] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.

[26] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2015.

[27] F. Chollet, "keras," https://github.com/fchollet/keras, 2015.

[28] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio, "Theano: new features and speed improvements," in *Proc. NIPS workshop Deep Learning and Unsupervised Feature Learning*, 2012.

[29] J. S. Baer, E. A. Wells, D. B. Rosengren, B. Hartzler, B. Beadnell, and C. Dunn, "Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors," *Journal of substance abuse treatment*, vol. 37, no. 2, p. 191, 2009.

[30] Z. E. Imel, M. Steyvers, and D. C. Atkins, "Computational psychotherapy research: Scaling up the evaluation of patient–provider interactions," *Psychotherapy*, vol. 52, no. 1, pp. 19–30, 2015.

[31] 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*, 2016, accepted.