INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Attention Networks for Modeling Behaviors in Addiction Counseling

James Gibson[1], *Doğan Can*[1], *Panayiotis Georgiou*[1], *David C. Atkins*[2], *Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA, USA
[2]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA
[1]`sail.usc.edu`, [2]`datkins@u.washington.edu`

Abstract

In psychotherapy interactions there are several desirable and undesirable behaviors that give insight into the efficacy of the counselor and the progress of the client. It is important to be able to identify when these target behaviors occur and what aspects of the interaction signal their occurrence. Manual observation and annotation of these behaviors is costly and time intensive. In this paper, we use long short term memory networks equipped with an attention mechanism to process transcripts of addiction counseling sessions and predict prominent counselor and client behaviors. We demonstrate that this approach gives competitive performance while also providing additional interpretability.

Index Terms: behavioral signal processing, recurrent neural networks, attention, word embedding, motivational interviews

1. Introduction

Human behaviors are extremely complex phenomena, occurring through several signals at a variety of rates, carrying a wealth of information. Psychotherapy is a semi-structured setting in which these behaviors are expressed, and elicited, for the purpose of improving clients' well-being. As these therapies are generally conversation based they provide an abundance of speech and language data, providing an opportunity for signal processing and machine learning researchers to contribute technological approaches to modeling and understanding these important interactions. Such approaches are increasingly becoming of interest, lending to the development of the fledgling field of behavioral signal processing [1].

In this work, we focus on a particular type of psychotherapy called motivational interviewing. Motivational interviewing (MI) is a client-centered, goal-oriented therapy, which focuses on resolving clients' ambivalence towards their problems in order to motivate behavior change [2]. This is achieved by the counselor exhibiting and eliciting certain desired behaviors, while avoiding certain undesired behaviors. Desirable counselor behaviors include reflective listening and asking open-ended questions, with the purpose of motivating the client to give reasons and make commitments towards behavior change. Undesirable counselor behaviors include confrontation and advising the client without their permission. Several studies suggest that the counselor's skill in adhering to the therapy goals influence client outcomes [3, 4].

Traditionally, psychotherapy interaction data is observed and annotated manually as part of a process referred to as behavioral coding in psychotherapy literature. This process is expensive and time-consuming, which prompts interest for automated solutions within the field. Several previous studies have focused on signal processing and machine learning of human behavior in psychotherapy interactions, including: modeling therapist session level empathy through language use [5] and prosodic cues [6], predicting utterance level therapist and client behaviors [7, 8, 9], and mapping between turn and session level behaviors [10]. Past work has largely relied on machine learning methods with limited interpretability. For feedback to human therapists, models with greater interpretability provide routes to describing why a predictive model chose a particular code or behavior.

Attention-based recurrent neural networks (RNNs) have gained popularity for their ability to select the appropriate context in sequence data while also giving interpretability towards which elements in the sequence lend most importance to the predictions made by the models. They've shown promise for a number of problems including image classification [11], describing multimedia content [12], document classification [13], and machine translation [14].

In this paper, we describe a corpus of motivational interviews and the behaviors with which they were annotated. There is considerable interest in how spoken language use relates to these behaviors [15, 16], thus we focus on lexical information for modeling in this paper. Specifically, we employ long short term memory (LSTM) networks [17] with an attention mechanism to consume the words in a given speaker turn to predict the behaviors that occur in that turn.

2. Motivational Interviewing Data

The corpus used for this work is comprised of motivational interviewing sessions collected from six independent clinical trials. All the studies focused on curbing various forms of addiction: four on alcohol abuse (ARC, ESPSB, ESB21, CTT), one on marijuana abuse (iCHAMP), and one on poly-drug abuse (HMCBI) [15, 18]. The CTT subset contains both real (N=76) and standardized (N=124) patients. Standardized patients are actors portraying clients struggling with addiction for the purpose of counselor training. All the other data subsets are comprised only of real patients. A subset of the sessions were human transcribed, and segmented at the turn level. Subsequently, 337 of the transcribed sessions were segmented at the utterance level and each utterance was annotated with behavioral labels according to the Motivational Interviewing Skill Code (MISC) manual [19].

There are 28 utterance level behaviors described in the MISC manual: 19 counselor and 9 client. We adopt the strategy used by Xiao et al., to reduce this to a set of 11 behavioral codes, 8 counselor and 3 client, by grouping the least frequently occurring codes into composite categories [8]. The non-grouped counselor categories are: facilitate (FA), giving information, (GI), simple reflection (RES), complex reflection (REC), closed questions (QUC), and open questions (QUO); all other counselor behavioral codes are grouped into MI adherent (MIA) and MI non-adherent (MIN) categories, i.e., behaviors which do or do not adhere to the aims of MI. Examples of non-adherent be-

`http://dx.doi.org/10.21437/Interspeech.2017-218`

haviors are confronting or warning the client about their addictive behaviors. The client behavior of follow/neutral is the only ungrouped client code; the rest of the client codes are grouped into positive (POS) and negative (NEG) categories, indicating client statements which are indicative of or counter to positive behavior changes. We give an overview of the behavior code grouping and counts in Table 1.

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

Group	MISC	Count
Counselor		
FA	Facilitate	15973
GI	Giving information	18120
QUC	Closed question	6343
QUO	Open question	5597
REC	Complex reflection	4053
RES	Simple reflection	6390
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 Reasons; Commitments; Taking steps; Other	6630
NEG	Sustain talk: negative Reasons; Commitments; Taking steps; Other	6218

The sessions are separated into training and testing sets according to an approximate 2:1 ratio (228 training and 109 testing). Some counselors appear in multiple sessions, so all the sessions from a given counselor are assigned to the same training/testing split. We give an overview of subject session, and word counts in Table 2.

Table 2: *Session, turn, and word counts in training/testing splits.*

Subject	Sessions	Turns	Words
Counselor	228/109	28.7K/13.9K	579K/248K
Client	228/109	28.6K/13.6K	563K/269K

3. Methodology

We consider each speaker turn from the sessions to be a sequence of words, $\mathbf{w} = \{w_1, w_2, \cdots, w_T\}$. Because turns can contain multiple utterances, they may also have multiple behavior labels assigned to them (e.g., a turn can contain both a question and a reflection). So the label, y, for a particular behavioral code corresponds to the presence or absence of that behavior in the turn. Note, this is a distinct approach from [8], where there is a one-to-one correspondence between utterances and behavioral labels. This choice is made to facilitate fully automated prediction, where segmenting by turn is a more well defined task than segmenting by utterance.

We first learn a word embedding vector representation of the words used by the counselors and clients in the motivational interviewing data. Each word, w, is represented by an M-dimensional vector, v (M=300). The word embedding vectors are input to the first layer of the network. The first layer is a single feed-forward layer used to fine tune the word vector representation:

$$x_{jt} = W_x v_{jt} + b_x, \tag{1}$$

for the t^{th} word in the j^{th} turn.

3.1. Attention-based LSTM

The fine tuned feature vectors are then fed into a forward LSTM, resulting in the hidden state vectors, h_{jt}, i.e.,

$$\{h_{j1}, h_{j2}, \cdots, h_{jT}\} = \text{LSTM}(\{x_{j1}, x_{j2}, \cdots, x_{jT}\}). \tag{2}$$

These hidden state vectors represent the words and their left context. Subsequently these representations are fed into the word attention mechanism.

The word attention mechanism consumes the hidden state vectors from the LSTM and feeds them to a single layer feedforward network that projects them into 1 dimension and applies the ‘tanh’ activation function, according to,

$$u_{jt} = \tanh(W_u h_{jt} + b_u). \tag{3}$$

These weights represent the estimated importance of the hidden vector at time, t. Subsequently, the weights are input to the softmax function for normalization across time, i.e.,

$$\alpha_{jt} = \frac{\exp(u_{jt})}{\sum_t \exp(u_{jt})}, \tag{4}$$

giving the relative importance of each hidden vector with respect to the other vectors from that turn. The hidden vectors are then combined via a weighted average according to their relative importance,

$$s_j = \sum_t \alpha_{jt} h_{jt}. \tag{5}$$

The final stage of the network is a single feedforward layer with a sigmoid activation to convert the weighted average vector into the probability of a particular behavioral being present in that turn, i.e.,

$$\hat{y}_j = \sigma(W_y s_j + b_y). \tag{6}$$

The network is trained to minimize the binary cross-entropy using the ADAM algorithm [20]. We show an overview of the proposed attention-based LSTM in Figure 1. For reference, we present the LSTM network without the attention mechanism, which in effect means $\alpha_{jt} = \frac{1}{T}, \forall t \in \{1, \cdots, T\}$.

3.2. Baseline

As a point of comparison, we use a feedforward neural network (FFNN) trained with word vectors using a bag-of-words approach. The fine tuned feature vectors, x_{jt}, are averaged over each turn to derive the turn vector representation, $r_j = \frac{1}{T}\sum_t x_{jt}$, which is then fed to a single feedforward layer with a sigmoid activation for prediction:

$$\overline{y}_j = \sigma(W_y r_j + b_y). \tag{7}$$

This network is also trained using binary cross entropy loss with the ADAM algorithm.

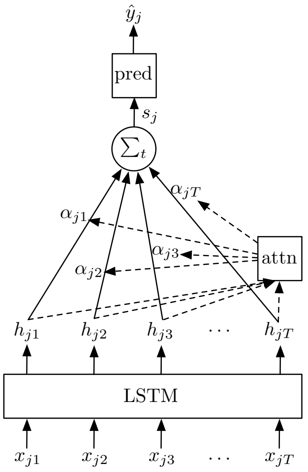


Figure 1: *Diagram of LSTM with attention*

Table 3: *MISC Prediction (F1-measure).*

code	FFNN	LSTM	LSTMa
Counselor			
FA	0.898	0.902	**0.904**
GI	0.697	0.733	**0.742**
QUC	0.550	0.704	**0.708**
QUO	0.667	0.816	**0.824**
REC	0.521	0.559	**0.562**
RES	0.443	0.533	**0.548**
MIA	0.533	**0.598**	0.592
MIN	0.261	0.259	**0.279**
Client			
FN	0.946	0.952	**0.953**
POS	0.452	0.463	**0.472**
NEG	0.406	**0.421**	0.419
avg	0.580	0.631	**0.637**

4. Experiments and Results

We use the word2vec software to train the word embeddings [21]. Keras, with Theano as the back end, is used to implement the neural networks [22, 23]. All hidden layers have the same dimension as the input feature vector. For regularization, 10% dropout is applied to all hidden layers. All turns are assumed to be 70 words in length, shorter turns are padded with zeros and longer turns are truncated (less than 5% of turns in the data required truncation). Ten percent of the training data is randomly chosen as a validation set. Training is performed with a maximum of 100 epochs and the training process is terminated early if the validation loss fails to improve after two consecutive epochs and only the weights from the epoch with the lowest validation loss are saved.

4.1. Behavioral Code Prediction

In Table 3, we show the F1-measure (the harmonic mean of precision and recall) for the baseline FFNN and LSTMs with and without attention (LSTMa and LSTM, respectively). For all behavioral codes, the attention-based LSTM gives better prediction than the baseline bag-of-words feedforward neural network baseline. This illustrates the importance of the relationship between words that is captured in the hidden states of the LSTMs. Adding attention to the traditional LSTM gives an improvement for all behavioral codes but MI adherent (counselor) and Negative (client). The highest relative improvements when using attention are given for RES (+2.8%) and MIN (+7.7%) for counselor categories and POS (+1.9%) for client. This may be a result of subjects using more salient words in the expression of these behavior categories.

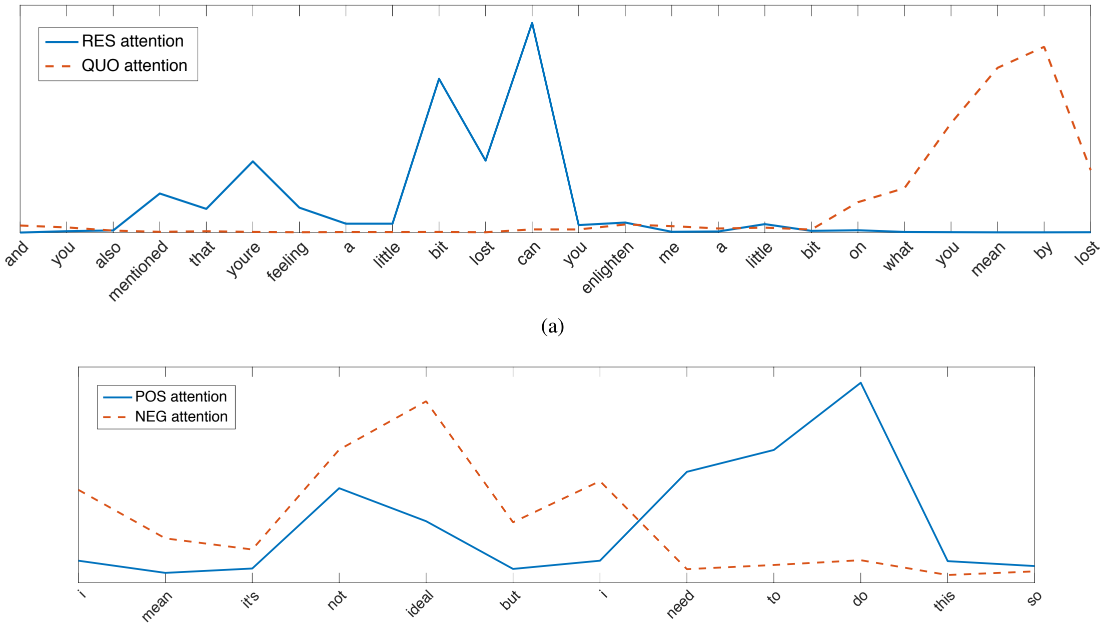
In Figures 2(a) and 2(b) we show examples of counselor and client turns and how the attention-based LSTM assigns attention over the turn. In the counselor turn (Figure 2(a)) there is a simple reflection followed by and open question. The RES attention mechanism focuses the earlier portions of the turn which corresponds to the reflective portion, “also mentioned that you’re feeling a little bit lost”, while the QUO attention mechanism ignores the early portion of the turn and focuses on the question at the end, “what do you mean by that”. Figure 2(b) shows an example of the POS and NEG attention for a client turn. This turn begins with a negative reason, “i mean it’s not ideal”, followed by a positive reason, “i need to do this”. The NEG attention focuses on the first part of the turn. The POS attention attributes some weight to the first part of the turn but applies much more attention to the positive portion of the turn.

4.2. Analysis of Attention Weights

In Table 4, we show the words which received the highest average attention for each behavioral code. Facilitate (FA) is an utterance by a counselor that is used to acknowledge what the client is saying and to indicate for them to keep going, hence words like ‘yeah’, ‘right’, and ‘good’ are attuned to for predicting this behavior. Giving information (GI) is an utterance where the counselor explains something or provides feedback. Interestingly, the word ‘mirror’ is identified as important because during many sessions the counselors explain to their clients that there is a one way mirror in the counseling room. Closed questions (QUC) are questions that prompt a specific fact as a response, whereas open questions (QUO) invite the client’s perspective. The word ‘what’ draws attention for both QUC and QUO as it is indicative of a question. Words in closed questions which receive attention derive from questions about specifics, e.g., “have you tried...” or “has anyone told you...”. In open questions, the words that receive attention come from questions such as “how does that fit...” or “what are your thoughts about...”.

Reflections are a key aspect of motivational interviewing. They are a way for the counselor to express understanding of the client’s perspective. Simple reflections (RES) are a restatement of something the client has said, while complex reflections (REC) add meaning or emphasis to something the client has expressed. Words that receive attention in simple reflections are specific words that were used in statements by the client such as ‘pot’ or ‘husband’. In complex reflections, words receiving attention are more likely to be less concrete in nature such as ‘scary’ or ‘felt’. Because MI adherent (MIA) and non-adherent (MIN) categories are groups of different behavior codes, they provide a challenge for attention. The MI adherent category is one where attention did not help the LSTM in prediction. This is essentially an ‘other’ category where any utterance that is not specifically one of the aforementioned behavioral codes but is also not a non-adherent behavior.

The client code follow/neutral (FN) is a catch all for utterances that are neither indicative of or counter to behavior



(a)

(b)

Figure 2: *(a) Example of attention for RES and QUO codes in counselor turn and (b) POS and NEG codes in client turn.*

change. Because of the nature of this code, attention is not particularly useful, however it does try to take advantage of some quirks in the data. For example, the word ‘island’ appears as an important word for this code because there are a few sessions in which long island iced teas are discussed at length. The POS and NEG categories combine behavioral codes relating to client statements about behavior change. The attention mechanism identifies words such as ‘control’ and ‘important’ which indicate commitment towards confronting problems as well as words such as ‘drive’ and ‘black’ which stem from statements about ramifications of substance use like driving under the influence and blacking out. Attention also identifies words that justify continued substance use such as ‘helps’ and ‘relax’ and words that indicate unwillingness to admit the existence of an issue such as ‘dunno’ and ‘problem’.

5. Conclusions and Future Work

In this paper, we presented an approach to predicting counselor and client behaviors in addiction counseling using attention-based LSTMs. We demonstrated that this approach out performs a bag-of-words approach and that attention boosts performance in some situations while providing additional insight into the data which is not afforded by non-attentional LSTMs. The attention weights help identify words that are salient in the speaker turns with respect to the behaviors present in these turns.

In the future, we would like to investigate using hierarchical attention networks [13] to perform prediction of both local (turn level) and global (session level) behaviors from the words used in the counseling sessions. Such an approach would allow for understanding not only the most important words in each turn but also the most important turns in each session. Additionally, we plan to evaluate attention-based LSTMs for transcripts derived from automatic speech recognition (ASR) in an end-to-end ‘sound to code’ system [24].

Table 4: *Words with the highest average attention for each behavior category.*

code	top words
Counselor	
FA	hmm, yeah, great, wow, cool, right, oh, good
GI	yup, include, mirror, thank, i’m, resources, yeah, participating
QUC	tried, questions, anything, anybody ever, what, heard, think
QUO	fit, does, what, happened thoughts, else, acted, do
REC	main, wanted, changes, scary, necessarily, could, hand, felt
RES	pot, husband, mentioned, month, sounds, thirty, earlier, hour
MIA	glad, cool, excellent, difficult, packet, we’ll, tough, great
MIN	help, these, give, could, try, able, keep, kids
Client	
FN	accurate, students, percent, assume, heavy, island, that’d, interesting
POS	control, drive, reason, us, important, black, quit, ten
NEG	relax, dunno, problem, helps smoked, eh, prefer, mostly

6. Acknowledgments

This work was supported by the NIH.

7. References

[1] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, 2012.

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

[3] J. Gaume, G. Gmel, M. Faouzi, and J.-B. Daeppen, "Counselor skill influences outcomes of brief motivational interventions," *Journal of Substance Abuse Treatment*, vol. 37, no. 2, pp. 151–159, 2009.

[4] J. McCambridge, M. Day, B. A. Thomas, and J. Strang, "Fidelity to motivational interviewing and subsequent cannabis cessation among adolescents," *Addictive Behaviors*, vol. 36, no. 7, pp. 749–754, 2011.

[5] B. Xiao, P. G. Georgiou, and S. Narayanan, "Analyzing the language of therapist empathy in motivational interview based psychotherapy," in *Asia-Pacific Signal and Information Processing Association*, 2012, pp. 1–4.

[6] B. Xiao, Z. E. Imel, D. C. Atkins, P. G. Georgiou, and S. S. Narayanan, "Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] D. Can, D. C. Atkins, and S. S. Narayanan, "A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

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

[9] 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.

[10] J. Gibson, D. Can, B. Xiao, Z. E. Imel, D. C. Atkins, P. Georgiou, and S. Narayanan, "A deep learning approach to modeling empathy in addiction counseling," *Commitment*, vol. 111, p. 21, 2016.

[11] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[12] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1875–1886, 2015.

[13] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of NAACL-HLT*, 2016, pp. 1480–1489.

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

[15] 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.

[16] S. P. Lord, E. Sheng, Z. E. Imel, J. Baer, and D. C. Atkins, "More than reflections: Empathy in motivational interviewing includes language style synchrony between therapist and client," *Behavior Therapy*, 2014.

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

[18] 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, pp. 191–202, 2009.

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

[20] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

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

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

[24] B. Xiao, Z. E. Imel, P. G. 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.