Journal of Substance Abuse Treatment 65 (2016) 43–50

Contents lists available at ScienceDirect

Journal of Substance Abuse Treatment

A Comparison of Natural Language Processing Methods for Automated Coding of Motivational Interviewing

Michael Tanana, M.Stat. [a,*], Kevin A. Hallgren, Ph.D. [b], Zac E. Imel, Ph.D. [c], David C. Atkins, Ph.D. [b], Vivek Srikumar, Ph.D. [d]

[a] University of UT, Department of Educational Psychology and Social Research Institute, 395 South 1500 East #111, United States
[b] University of Washington, Department of Psychiatry and Behavioral Sciences, Box 354944, Seattle, WA 98195-4944, United States
[c] University of Utah, Department of Psychology, 1705 Campus Center Drive, Room 327, Salt Lake City, UT United States
[d] University of Utah, School of Computing, 50 S. Central Campus Drive Room 3190, Salt Lake City, UT, United States

ARTICLE INFO

Article history:
Received 1 August 2015
Received in revised form 13 January 2016
Accepted 23 January 2016

Keywords:
Behavioral coding
Discrete sentence feature model
Machine learning
Motivational interviewing
Natural language processing
Recursive neural network
Treatment integrity

ABSTRACT

Motivational interviewing (MI) is an efficacious treatment for substance use disorders and other problem behaviors. Studies on MI fidelity and mechanisms of change typically use human raters to code therapy sessions, which requires considerable time, training, and financial costs. Natural language processing techniques have recently been utilized for coding MI sessions using machine learning techniques, rather than human coders, and preliminary results have suggested these methods hold promise. The current study extends this previous work by introducing two natural language processing models for automatically coding MI sessions via computer. The two models differ in the way they semantically represent session content, utilizing either 1) simple discrete sentence features (DSF model) and 2) more complex recursive neural networks (RNN model). Utterance- and session-level predictions from these models were compared to ratings provided by human coders using a large sample of MI sessions ($N = 341$ sessions; 78,977 clinician and client talk turns) from 6 MI studies. Results show that the DSF model generally had slightly better performance compared to the RNN model. The DSF model had "good" or higher utterance-level agreement with human coders (Cohen's kappa > 0.60) for open and closed questions, affirm, giving information, and follow/neutral (all therapist codes); considerably higher agreement was obtained for session-level indices, and many estimates were competitive with human-to-human agreement. However, there was poor agreement for client change talk, client sustain talk, and therapist MI-inconsistent behaviors. Natural language processing methods provide accurate representations of human derived behavioral codes and could offer substantial improvements to the efficiency and scale in which MI mechanisms of change research and fidelity monitoring are conducted.

1. Introduction

Motivational interviewing (MI; Miller & Rollnick, 2012) is a counseling style that advises specific interpersonal and linguistic strategies for helping clients highlight and resolve ambivalence about behavioral change. MI theory is based strongly on the interrelationships between clinician speech, client speech, and client long-term behavioral change (e.g., reduced substance use; Miller & Rollnick, 2012; Miller & Rose, 2009). In consequence, research and training in MI have relied strongly on behavioral coding to assess and understand clinicians' and clients' within-session verbal behavior.

Behavioral ratings provide critical data for research and training purposes. For example, coding data are often used clinically to assess MI fidelity (e.g., based on the amount of MI-consistent and MI-inconsistent clinician behaviors; De Jonge, Schippers, & Schaap, 2005; Moyers, Martin, Catley, Harris, & Ahluwalia, 2003) and to provide ongoing consultation and feedback to maintain high-quality MI practice (Schwalbe, Oh, & Zweben, 2014). In research, behavioral coding is increasingly used to monitor internal validity (i.e., adherence to treatment protocols) for clinical trials, to evaluate the effectiveness of MI training programs (e.g., by monitoring changes in clinician fidelity in response to training; Schwalbe et al., 2014), to understand the temporal relationships between clinician and client behaviors (e.g., through sequential analysis of clinician and client codes), and to model how specific in-session behaviors predict out-of-session behavioral change (e.g., clinician and client language predicting reductions in substance use; Bertholet, Faouzi, Gmel, Gaume, & Daeppen, 2010; Moyers, Martin, Houck, Christopher, & Tonigan, 2009).

Thus, behavioral coding has become an essential tool; however, the methods used in this work are often time- and cost-intensive (Forsberg, Berman, Kallmén, Hermansson, & Helgason, 2008). For example, coding

* Corresponding author.
E-mail addresses: Michael.Tanana@utah.edu (M. Tanana), khallgre@uw.edu (K.A. Hallgren), zac.imel@utah.edu (Z.E. Imel), datkins@uw.edu (D.C. Atkins), svivek@cs.utah.edu (V. Srikumar).

http://dx.doi.org/10.1016/j.jsat.2016.01.006
0740-5472/

studies often require formal coder training for a substantial period prior to coding and ongoing reliability assessments for the duration of the study period. A single 50-minute session can have 12,000–15,000 words contained within several hundred utterances, and can take between 20-minutes to several hours to code in full, depending on the complexity of the coding system. A single study may consist of hundreds of counseling sessions, ultimately requiring substantial resources in the form of time, money, and personnel. These limitations impose substantial barriers for MI training and research efforts. For example, although the vast majority of substance abuse treatment facilities purport using MI (Substance Abuse and Mental Health Services Administration, 2014), only an extremely small number of MI sessions from these settings are actually coded and studied for fidelity, due in large part to the costs required for behavioral coding. Many clinicians who receive MI training get little or no subsequent coding-based evaluation or feedback on their MI delivery (Miller, Sorensen, Selzer, & Brigham, 2006), even though this feedback has been shown to facilitate improvement in MI skill (Schwalbe et al., 2014). Likewise, the burden of coding has led most MI coding studies to have small-to-moderate sample sizes (e.g., ranging from 30 to 195 in a recent meta-analysis of MI mechanism of change studies; Magill et al., 2014), limiting the statistical power of such studies. Compared to the scope of research on MI outcomes (e.g., 204,415 MI sessions reported in a meta-analysis by Lundahl, Kunz, Brownell, Tollefson, & Burke, 2010), very few MI sessions are actually coded and used in mechanisms of change research (e.g., 783 sessions in a meta-analysis by Magill et al., 2014). The use of small samples considerably limits progress in MI research, for example, limiting finer-grained analyses to understand low-frequency behaviors (e.g., confrontation) or specific behavioral codes (e.g., complex reflections) rather than broader collapsed categories (e.g., MI-consistent behaviors), and often limiting the testing of various interactions (e.g., moderated mediation models) based on various contextual variables (e.g., patient, provider, setting).

While the importance of behavioral coding is clear, its utility is restricted by the time and cost limitations of manual coding, thus necessitating the development of automated coding tools. Natural language processing, a subfield of computer science that focuses on understanding and modeling human language, has been increasingly used in recent decades in a variety of applications. For example, these methods have been applied to understand word relationships in Wikipedia (Mikolov, Corrado, Chen, & Dean, 2013), automatically extract the sentiment of movie reviews (Pang & Lee, 2005), evaluate the emotional content of Facebook status updates (Kramer, Guillory, & Hancock, 2014), categorize articles in the New York Times (Newman, Chemudugunta, Smyth, & Steyvers, 2006), answer biology text book questions (Berant et al., 2014), and examine themes in poetry (Kao & Jurafsky, 2012). Excellent and approachable introductory overviews of natural language processing are available in Hirschberg and Manning (2015) and Jurafsky and Martin (2008).

Modern natural language processing methods have only recently been applied to in-session language in psychotherapy research (e.g., see Maskit, Bucci, & Murphy, 2015; Rouhizadeh, Prud'hommeaux, Roark, & van S., 2013). In one of the first tests of computer-predicted coding in MI, Can et al. (in press) used maximum entropy Markov models to predict clinician reflections based on the occurrence of words and phrases within utterances (e.g., "it sounds like") as well as the amount of overlap of words and phrases between clinician and client speech. They obtained good accuracy for predicting reflections (positive predictive value = 0.73, sensitivity = 0.93), however, this model only predicted clinician reflections but not other types of clinician or client speech and was limited to 57 sessions. Atkins, Steyvers, Imel, and Smyth (2014) used a different technique called topic modeling to predict a larger set of clinician and client behavioral codes from 148 MI sessions. These topic models were latent variable mixture models for text, in which text documents were seen as a mixture of underlying "topics" and each topic is a probability distribution over words. Atkins et al. used a labeled topic model in which the model learns topics that are specifically related to behavioral codes and other background topics. (For an overview of topic models applied to psychotherapy transcripts, see Atkins et al., 2012.) The topic models provided good predictive performance for many clinician codes, such as open and closed questions and simple and complex reflections (Cohen's kappa all > 0.50), but struggled to accurately predict other codes, particularly client change talk and sustain talk (Cohen's kappa < 0.25). Several limitations may have contributed to the model's difficulty with some codes; for example, topic models use a "bag of words" approach, where the functional relationships and ordering between words are ignored. Imel, Steyvers, and Atkins (2015) used similar topic modeling techniques with a larger corpus of 1,553 psychotherapy sessions (including, but not limited to MI and substance abuse treatment) and found that topic models were able to identify various treatment-related topics (e.g., emotions, relationships) and were able to discriminate different types of treatments (e.g., psychodynamic, humanistic, and medication management therapies).

Overall, these initial findings suggest that these methods hold promise to speed up and simplify the process of behavioral coding; however, significant research is still needed to develop and explore which types of natural language processing models are the most suitable for predicting behavioral coding data in the context of MI practice and research. It is likely that capturing the complexities of MI in-session behavior will require ongoing comparisons of different natural language processing models that quantify different aspects of the linguistic exchange.

The present study aims to build on this line of work by testing and comparing the results from two natural language processing methods for predicting behavioral codes from transcripts of MI sessions. The two methods, introduced and described below, differ substantially in terms of their complexity and their conceptualization of language and linguistic structure. Moreover, they are fundamentally different models relative to topic models or maximum entropy models tested in previous research (Atkins et al., 2014; Can et al., in press; Imel et al., 2015). The most important difference between the models tested in this study and topic models is that these models attempt to incorporate the linguistic structure of a sentence beyond the mere presence of words and phrases. In addition to introducing and testing each model, the present study also comprises the first head-to-head comparison of two different natural language processing methods for coding MI sessions, which will help MI research by honing in on the best methods for modeling clinician and client language in MI sessions. When paired with existing (and improving) methods for automated speech recognition (i.e., converting voice to text), the methods for processing natural language tested here (i.e., converting text to theoretically-relevant MI codes) could fundamentally change the manner and scale by which MI sessions can be coded.

2. Material and Methods

2.1. Data Source

We evaluated utterances and behavioral codes from 341 psychotherapy sessions in 6 MI clinical trials, including the 148 sessions (5 studies) reported in a previous test of topic models for computer-based MI coding (Atkins et al., 2014). The original studies tested the effectiveness of MI for substance use within a large, public safety net hospital (Roy-Byrne et al., 2014; 70 sessions), the efficacy of training clinicians in MI (Baer et al., 2009; 195 sessions), and the efficacy of MI in reducing college student drinking (Lee et al., 2013, 2014; Neighbors et al., 2012; Tollison et al., 2008; 76 sessions). Combined, these sessions contained approximately 1.7 million words, 175,000 utterances, and 79,000 talk turns.

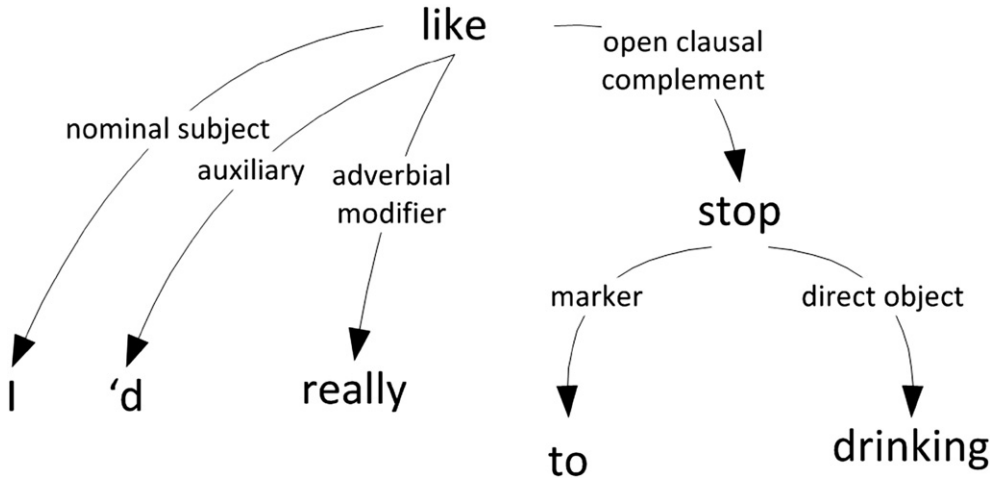

Fig. 1. Dependency tree for a hypothetical change talk statement.

2.2. Behavioral Coding Measure

Human raters coded all sessions using a modified version of the Motivational Interviewing Skill Code (MISC version 2.1; Miller, Moyers, Ernst, & Amrhein, 2008). Raters assigned a single, categorical behavioral code to each client and clinician utterance (i.e., single idea or talk turn; Lord et al., 2014). Unlike the original MISC 2.1 where utterances are parsed and then coded in separate coding passes, utterances were parsed and coded at the same time within a single pass. Clinician behavior codes included several categories that are broadly categorized as MI-consistent (e.g., open questions, complex reflections), MI-inconsistent (e.g., confront, advise without permission), or MI-neutral (e.g., giving information); client codes are either change talk (i.e., language that favors changing substance use), sustain talk (i.e., language that favors maintaining substance use), or follow/neutral (i.e., neither change talk nor sustain talk). In the present study, only MISC behavioral codes and session-level MISC summary indices (e.g., percent open questions) were predicted; global ratings (e.g., overall empathy or MI spirit for the entire session) were not predicted[1] (however, for examples, see Can, Atkins, & Narayanan, 2015; Imel et al., 2014, and Xiao, Georgiou, Imel, Atkins, & Narayanan, 2013, Xiao et al., 2014, Xiao, Imel, Atkins, Georgiou, & Narayanan, 2015).

2.3. Natural Language Processing Models

Two natural language processing models were tested, both predicting utterance-level MISC codes from transcript text. In the present application, human raters parsed session transcripts into utterances and coded them as clinician or client speech. These utterances were used as input to the models. Both models used machine learning, a broad class of methods in which a model "learns" to predict a particular outcome (e.g., MISC codes) from a given set of inputs (e.g., transcript text). The prediction models were repeatedly updated during a process called training, where the model-based predictions were compared to their corresponding known outcomes (e.g., human-rated MISC codes). During the training period, incorrect predictions cause the model to adjust the manner in which it predicts the outcome, continuously improving its predictive performance in a strictly data-driven approach. After the training period, the models underwent a testing period where their predictive performance was empirically tested with new data that were not seen during the training period.

Both models described below included *dependency trees* as input variables (De Marneffe, MacCartney, & Manning, 2006). Dependency trees model the grammatical structure of natural language by linking words into a hierarchical structure that accounts for the ordering and functional relationships between words (e.g., subject–verb, verb–object relationships). The use of dependency trees as inputs allows the model to include information about the specific words, as well as the relationships among the words as inputs into the prediction models even if they are located in distant parts of the sentence. An example dependency tree is shown in Fig. 1, which displays a hypothetical client change talk utterance and the dependencies between words within the utterance. The dependency tree maps the functional relationships between words (e.g., "I" being the nominal subject of the verb "like"), allowing for more preservation of the overarching meaning of the utterance than if these functional relationships and/or the word ordering was ignored (e.g., although the words "like" and "drinking" both appear in the utterance, the relationship between them is, importantly, modeled through the words "to stop"). The dependency trees used in this study were automatically generated using the Stanford Parser (version 3.5.2, Manning et al., 2014), a free and commonly used natural language processing tool. The Stanford Parser first labels speech parts for individual words (e.g., noun–plural) then builds dependency trees to map relationships between words. The tool was developed from a number of large, written English language text corpora and utilizes machine learning algorithms to identify word annotations and dependencies.

The methods tested here differ from other research predictive models based on topic modeling (also called Latent Dirichlet Allocation; e.g., Atkins et al., 2014; Imel et al., 2015; Schwartz et al., 2013) and word-count methods using predefined keywords (e.g., Linguistic Inquiry and Word Count [LIWC]; Tausczik & Pennebaker, 2010) in a few key ways. First, unlike topic models, the present methods do not identify latent "topics" based on patterns of word co-occurrences. These topics can be used to represent themes or ideas communicated within utterances that can then be used to predict behavioral codes. In addition, unlike LIWC and other dictionary-based word count methods, the present study makes no a priori assumptions about specific keywords mapping onto specific constructs; instead, machine learning algorithms learn to predict codes from sentence features in a data-driven manner rather than a theory-driven one. Unlike both topic models and LIWC methods, the models described below incorporate dependency tree information about the functional relationships between words by using the information from the Stanford Parser. We now describe these methods here in more detail, and refer readers to the technical supplement for more information.

[1] In contrast to behavioral codes tested here, global ratings are Likert-type scale ratings based on behavior across the full interview; automated prediction of these codes therefore require alternative analytic procedures to those tested here.

2.3.1. Discrete Sentence Feature (DSF) Model

The first prediction model utilized N-grams and dependency parse tree data as inputs to the prediction model. N-gram inputs indicated whether specific words (unigrams, e.g., "stop"), sequential word pairs (bigrams, e.g., "stop drinking"), or word triplets (trigrams, e.g., "to stop drinking") occurred within an utterance; dependency parse tree data input variables indicated whether specific word co-dependencies were present in the model (e.g., the nominal-subject relationship between "I" and "like"). We compiled a full set of discrete sentence features that included all N-grams and dependency parse tree relations occurring at least twice within the training dataset. The relative likelihood of an utterance falling into a specific MISC category was predicted by the presence or absence of each discrete sentence feature within each utterance using a large multinomial regression model with tens of thousands of discrete sentence features (i.e., a binary indicator of the presence or absence of each word codependency or N-gram) as input variables.[2] For a more technically detailed introduction to N-gram models, see Jurafsky and Martin (2008).

2.3.2. Recursive Neural Network (RNN) Model

The second prediction model was based on *recursive neural networks* (RNNs; Socher, 2014). The RNN model used the same dependency tree nodes as the DSF model, but through a different internal structure for representing the information contained within text. Whereas the DSF model focuses on each individual word, short phrase, or dependency feature regardless of its context, the RNN model uses a multivariate "embedding" to represent each word. Each word in an utterance was initially assigned a vector of 50 numeric values[3] derived from the Stanford GloVe Project (Pennington, Socher & Manning, 2014). This numeric vector can be thought of as representing the semantics of each word within a 50-dimensional latent space. The positioning represented by this numeric vector modeled both semantic similarities between words (e.g., "king" and "queen" would be located near each other in the 50 dimensional space; Mikolov, Yih, & Zweig, 2013) and semantic differences between words (e.g., the difference between "king" and "queen" would be approximately equal to the difference between "man" and "woman").

The numeric vectors for each word were then iteratively combined into single numeric vectors at each edge of the dependency tree. For example, the numerical vectors representing "stop" and "drinking" would be combined to create a single numerical representation of the phrase "stop drinking." After this process was applied to each edge in the tree, all of the vectors were combined to create a single vector representation of the full utterance.[4] Here, the final representation (technically, neural layer) can be thought of as a set of logistic regressions, where several sets of continuous inputs (i.e., the numeric vectors for each type of word codependency) were combined to create a single set of outputs (i.e., a single numeric vector) with values constrained between zero and one (or negative one and positive one). The single output vector values were then used as input to a multinomial regression model to predict MISC codes.

2.4. Model Training

All models were trained to maximize their prediction of MISC codes using a set of 232 sessions (70% of full sample) that were randomly selected to be used only for training (the number of sessions was balanced across the parent studies). This training set was also utilized when empirically evaluating model features during a preliminary development phase (e.g., identifying the optimal size for numeric vector representations of words and identifying optimal learning rates). Once the final DSF and RNN models were developed and trained, a separate testing set of 109 sessions was used to test the prediction performance of both models. This separation of training data and test data is standard in machine learning and provides a gauge of a model's true prediction via its performance on sessions not previously "seen" by the model. All results below are reported on the test dataset.

2.5. Assessment of Model Performance

Model performance was assessed by comparing computer-predicted codes to human codes, both at the utterance level (i.e., utterance-by-utterance agreement) and at the session level for total counts of MISC codes and for summary scores commonly used in clinical and research contexts (e.g., percentage of open questions). Utterance-level agreement was computed using Cohen's kappa, which estimates inter-rater agreement while correcting for chance agreement and is commonly used in MI research (additional computational details and alternative agreement indices used in machine learning research are included in supplementary materials). Model performance of session-level code counts and MI summary scores was assessed using two-way, absolute-agreement, single-measures intraclass correlations (ICCs). Human inter-rater agreement was examined for a subset of sessions ($n = 63$) and served as an important reference point for comparing the tested models.

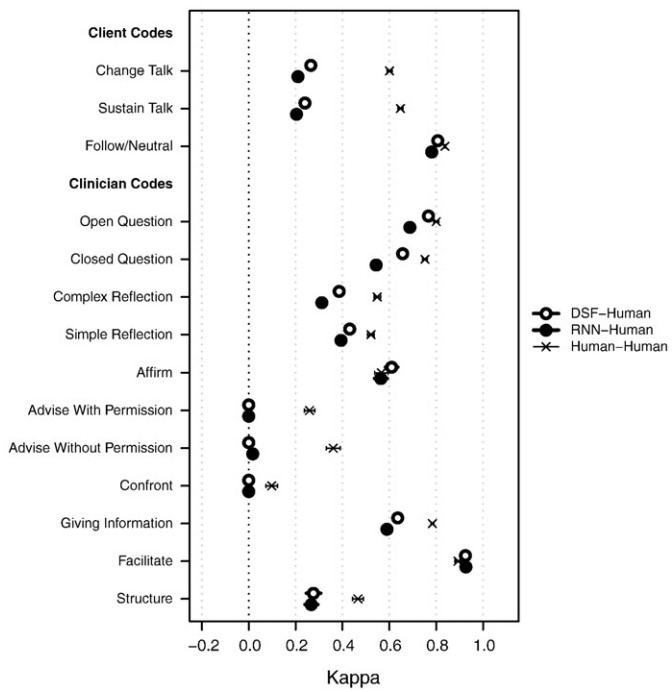

Fig. 2. Utterance-level Cohen's kappa comparing agreement for DSF–human, RNN–human, and human–human agreement.

[2] The goal of training was to minimize the squared errors in the predictions of the codes. Both models were trained using stochastic gradient descent, an iterative and relatively fast training method for datasets. The models were trained by predicting codes for one randomly-selected case at a time, and corrections to prediction parameters were incorporated each time a prediction error was made. Additional technical details are provided in supplementary materials.

[3] A 50-dimensional vector was selected based on optimal performance using 10-fold cross-validation with the training data.

[4] Vectors were combined by multiplying them through a set of weight matrices, with different weight matrices for each type of word-dependency. The matrices were initialized with random values at the start of the training period, then pre-trained on an unlabeled dataset. The weight matrices were then updated during the training period using backpropagation through structure (Goller & Kuchler, 1996), such that prediction errors created updates to the weight matrices to improve prediction accuracy. Specifically, the models used ada-grad with the diagonal variant (Duchi, Hazan, & Singer, 2011), which has faster learning for new words and slower learning for seen words. The 50 values in the final, top-level vector, which represented the whole utterance, were then used as input into a multinomial logistic regression model that predicted specific MISC codes. Additional technical details are provided in supplementary materials.

3. Results

3.1. Utterance-Level Prediction

The performance of DSF and RNN models for correctly predicting human-based, utterance-by-utterance codes is shown in Fig. 2. Performance for both models varied considerably by code, with both performing significantly better than chance for all codes except advise with permission, advise without permission, and confront, which all occurred infrequently across the full sample of sessions (<1% of utterances) and had the lowest human-to-human inter-rater agreement. The low frequency of these utterances is known to create difficulties in prediction modeling, and the poor performance of these codes is consistent with other research (e.g., Atkins et al., 2014).

The DSF and RNN models both tended to have high agreement with human coders for open and closed questions, facilitate, giving information, affirm and follow/neutral (all kappas > 0.50), with modestly lower utterance-level agreement for simple and complex reflections (all kappas between 0.30 and 0.50), and much lower agreement for client change talk and sustain talk (all kappas between 0.20 and 0.30).

For all codes, the DSF model performed equally well or better than the RNN model. In particular, the DSF model outperformed the RNN model for open and closed questions, complex reflections, and change talk, which had differences in kappas ranging from .055 to .113.

Examples of correct and incorrect model-classified utterances are presented in Fig. 3. The top model-predicted codes from both models are shown for each code and their corresponding scores. For example, in panel A, both models (correctly) predicted change talk as the most likely code for the corresponding text, with follow/neutral as the next most likely code and sustain talk as the least likely code. However, in panel B, the DSF model (correctly) predicted change talk as the most likely code but the RNN model (incorrectly) predicted follow/neutral as the most likely code. Similar examples are presented in panels C and D with therapist language. These examples illustrate how DSF and RNN models provide continuous scores reflecting the model-implied likelihood of the text corresponding to different MISC codes. In all cases, the code with the highest score is identified as the most likely MISC code that corresponds with the text.

3.2. Session-Level Prediction

One of the most common clinical uses of behavioral coding is to obtain session-level tallies of behavioral codes and session-level MI adherence summaries. The session-level results show stronger concordance between the two prediction models and human-based coding. Summing codes within sessions (Fig. 4, top section) resulted in agreement indices that were often considerably higher than utterance-level agreement. Using Cicchetti's (1994) guidelines for interpreting inter-rater agreement, affirm, facilitate, giving information, follow/neutral, simple reflections, and open and closed questions all were in the excellent range (ICC > 0.75); sustain talk was in the good range (0.60 < ICC < 0.75); and confront, structure, and advise with and without permission were in the poor range (ICC < 0.40). Complex reflections and change talk were in the good range for the DSF model but only were fair (0.40 < ICC < 0.60) for the RNN model.

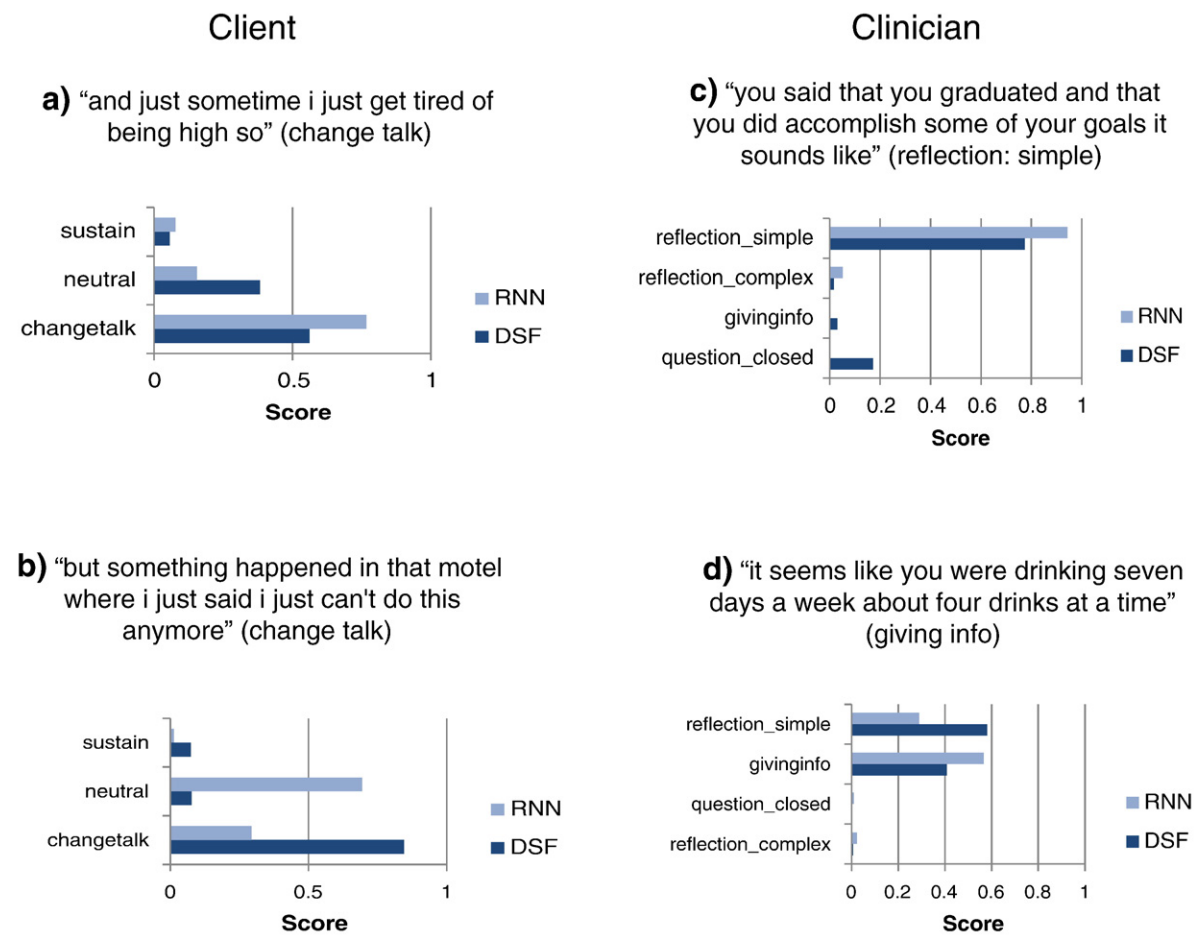

Fig. 3. Examples of correct and incorrect model classification (top predictions only). The top panels (a and c) show examples where the DSF and RNN models both classified utterances correctly. In the bottom panels (b and d), only one model correctly classified the utterance, the DSF (b) or the RNN (d). Both models showed varying degrees of uncertainty in classifying utterances, as indicated by non-zero scores for multiple codes.

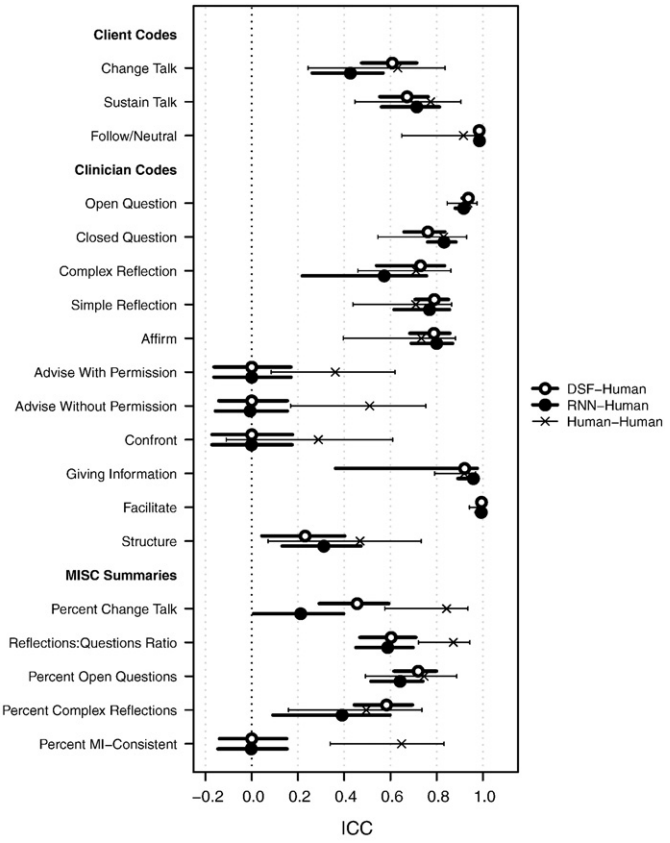

Fig. 4. Session-level ICCs comparing agreement for DSF–human, RNN–human, and human–human agreement.

Model-predicted session-level MI-adherence summaries are presented in the bottom section of Fig. 4. Overall, the DSF model performed better than the RNN, the former yielding agreement estimates that were in the fair-to-good range for all codes except the percentage of MI-consistent clinician behaviors, which was poor for both models.

4. Discussion

In the current work we have tested two new models for predicting MISC codes in MI sessions, recursive neural network and discrete sentence feature models, and compared their prediction accuracy with human raters. Both models attempt to incorporate syntactic and semantic information in a higher-level manner than typical word counting techniques. This study adds to a growing body of work showing how natural language processing methods can predict behavioral codes in psychotherapy sessions (Atkins et al., 2014; Can, Georgiou, Atkins, & Narayanan, 2012; Can et al., 2015). In addition to testing new models, we have also increased the number of sessions previously used in this line of work from 148 sessions (Atkins et al., 2014) to 341.

This study found that both models predicted many of the codes well at the utterance level and at the session level. For most codes, session-level agreement was higher than utterance level agreement. Many of the model-created predictions were comparable to human reliability estimates even at the utterance level (follow-neutral, open and closed questions, simple reflections, affirm and facilitate). However, there were some codes that both models had difficulty with, and this was particularly true with rare codes such as advise with or without permission and confront.

Model predictions for change talk and sustain talk had agreement that was comparable to human ratings at the session level, but had lower agreement at the utterance level. Part of the reason for this is that many of these statements are difficult even for humans to classify, and coders often must rely on information that is subtle or dependent on earlier speech outside of the particular utterance being coded. This type of contextual information was not included in the machine learning techniques tested here. In addition, change and sustain talk are thought to reflect specific aspects of motivation (e.g., attitudes toward using substances), which inherently involve greater semantic complexity to distinguish them from other types of speech. In contrast, other MI codes are based more strongly on specific semantic sentence structures or can be more easily identified through specific keywords (e.g., closed questions), which likely contributed to their better performance in the present study.

Overall, in this study we found that the DSF model performed slightly better than the RNN model. There are several possible reasons for this. For example, although the dataset in the present study was larger than in previous work, it may have still been too small to adequately estimate the large number of parameters that the RNN utilizes. In addition, the RNN model relies on word vectors that were originally trained on more formal written sources (i.e., Wikipedia), not conversational data; thus, RNN models may perform better if initial word vectors can be obtained through sources based on conversations or psychotherapy sessions. Nonetheless, the results obtained here suggest that the use of discrete sentence features (e.g., unigrams, bigrams, trigrams, and word dependency features) are likely sufficient for accurately predicting many MISC codes.

In comparison to previous work predicting multiple MISC codes (Atkins et al., 2014), the results here had higher utterance-level kappas for affirm, change talk, giving information, closed and open questions, simple reflections, and sustain talk, but poorer agreement for complex reflections. However, care should be taken when comparing the results of the present study to previous work, as the current sample size was considerably larger and more heterogeneous. Thus, any differences between the current study and previous studies cannot be attributed solely to model performance, and it is likely that some of the differences here are due to having a different and substantially larger training dataset.

One limitation of the present study is that codes were predicted only by the text within a particular utterance. Future work could attempt to overcome this limitation by including other contextual information from previous utterances or incorporating acoustic data into the predictions to improve model performance. For example, prosody (i.e., the way that a person varies their tone) has a major impact on the way that humans perceive the differences between questions and reflections or other types of statements (Cutler, Dahan, & van Donselaar, 1997). It is possible that these kinds of features could be integrated into more complex models of language in MI in the future (Sridhar, Bangalore, & Narayanan, 2008). Surprisingly, in our initial tests for the present study, we were unable to improve our model performance including information from the previous or subsequent utterance. Psychotherapy generally, and MI specifically, involve a dynamic exchange between two human beings, but we have not yet found effective ways to capture the dyadic nature of this process. As methods continue to improve for modeling utterances within the context of larger conversations, we hope that MI prediction models will likewise grow in sophistication and prediction accuracy.

Another limitation of the clinical trial datasets used in the present study is that MI-inconsistent behaviors occurred relatively infrequently. As a result, both models had difficulty learning to identify these codes, and even the human raters often disagreed about these rare codes. This is likely the result of many of these datasets coming from structured clinical trials, which usually have higher clinician adherence than real world settings. Thus, adding more naturalistic, non-clinical trial sessions to this dataset would likely improve prediction accuracy for these types of utterances. The models examined here typically perform the best with very large datasets, and even though the dataset used here is very large for its kind, these and other machine learning-based models would likely benefit from even larger training datasets.

To allow readers to examine the performance of these models beyond the examples included here, we have created an online prediction model, a beta version of which can be viewed at (http://compmh.ed.utah.edu/psychtest/). With this interactive tool, users have the ability to enter sample client and clinician utterances, obtain DSF- and RNN-predicted codes, and provide feedback for incorrect predictions. It is our hope that with wider use, our models might benefit from the feedback of outside experts and practitioners and learn to correctly identify statements beyond the ones currently available in our database. For example, there were very few confront codes in our dataset; if MI practitioners tested and corrected the system with many examples of confront, our models may improve at predicting these statements.

5. Conclusion

While natural language processing methods tested here have room for improvement, they suggest that there is promise for a future in which clinician adherence and within-session mechanisms of change could be monitored efficiently, accurately, and on a substantially larger scale than is currently feasible with human coding. With further improvements, and in combination with automatic speech recognition, it is not unreasonable to imagine scenarios where in-session audio could be directly uploaded to a computer, then automatically transcribed and coded in real time. Thus, clinical supervisors or clinicians themselves could review adherence for *every* session that their clinicians conduct. Such technology could allow clinical supervision to be an ongoing part of practice, instead of a rare and often random occurrence.

Acknowledgements

This current research was supported by grants from the National Institute on Alcoholism and Alcohol Abuse (NIAAA, R01AA018673, K02AA023814, T32AA007455) and the National Institute on Drug Abuse (NIDA, R34DA034860). The original randomized trials were supported by grants from NIAAA (R01AA016979, R01AA014741) and NIDA (R01DA025833, R01DA026014). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors acknowledge the contributions of our MISC rating team for their efforts in generating the human-based MISC codes. We are grateful for the use of EMULAB (White et al., 2002) for computational resources used to train the models in this paper and to Padhraic Smyth for a number of helpful comments he made throughout the development of this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jsat.2016.01.006.

References

Atkins, D. C., Rubin, T. N., Steyvers, M., Doeden, M. A., Baucom, B. R., & Christensen, A. (2012). Topic models: A novel method for modeling couple and family text data. *Journal of Family Psychology*, *26*, 816–827.

Atkins, D. C., Steyvers, M., Imel, Z. E., & Smyth, P. (2014). Scaling up the evaluation of psychotherapy: Evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, *9*, 1–11.

Baer, J. S., Wells, E. A., Rosengren, D. B., Hartzler, B., Beadnell, B., & Dunn, C. (2009). Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors. *Journal of Substance Abuse Treatment*, *37*, 191–202.

Berant, J., Srikumar, V., Chen, P., Huang, B., Manning, C. D., Vander Linden, A., & Harding, B. (2014). Modeling biological processes for reading comprehension. *Proceedings of EMNLP* (pp. 1499–1510).

Bertholet, N., Faouzi, M., Gmel, G., Gaume, J., & Daeppen, J. B. (2010). Change talk sequence during brief motivational intervention, towards or away from drinking. *Addiction*, *105*, 2106–2112.

Can, D., Atkins, D. C., & Narayanan, S. S. (2015). *A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations.* Paper to be presented at *Interspeech-2015*.

Can, D., Georgiou, P. G., Atkins, D. C., & Narayanan, S. (2012). A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features. *INTERSPEECH, 13th Annual Conference of the International Speech Communication Association* (pp. 2251–2254).

Can, D., Marin, R. A., Georgiou, P. G., Imel, Z. E., Atkins, D. C., Narayanan, S. S. (2016) "It sounds like…": A natural language processing approach to detecting counselor reflections in motivational interviewing. *Journal of Counseling Psychology* (Epub ahead of print).

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290.

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, *40*(Pt 2), 141–201.

De Jonge, J. M., Schippers, G. M., & Schaap, C. P. D. R. (2005). The Motivational Interviewing Skill Code: Reliability and a critical appraisal. *Behavioural and Cognitive Psychotherapy*, *33*, 285–298.

De Marneffe, M. -C., MacCartney, B., & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 449–454).

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning*, *1–40*.

Forsberg, L., Berman, A. H., Kallmén, H., Hermansson, U., & Helgason, A. R. (2008). A test of the validity of the motivational interviewing treatment integrity code. *Cognitive Behaviour Therapy*, *37*, 183–191.

Goller, C., & Kuchler, A. (1996). Learning task-dependent distributed representations by backpropagation through structure. *Proceedings of International Conference on Neural Networks (ICNN'96)* (pp. 1). http://dx.doi.org/10.1109/ICNN.1996.548916.

Hirschberg, J., & Manning, C. (2015). Advances in natural language processing. *Science*, 261–266.

Imel, Z. E., Barco, J. S., Brown, H. J., Baucom, B. R., Kircher, J. C., Baer, J. S., & Atkins, D. C. (2014). The association of therapist empathy and synchrony in vocally encoded arousal. *Journal of Counseling Psychology*, *61*, 146–153. http://dx.doi.org/10.1037/a0034943.

Imel, Z. E., Steyvers, M., & Atkins, D. C. (2015). Psychotherapy computational psychotherapy research: Scaling up the evaluation of patient – provider interactions. *Psychotherapy*, *52*, 19–30.

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, *104*, . http://dx.doi.org/10.1016/S0065-230X(09)04001-9.

Kao, J., & Jurafsky, D. (2012). A computational analysis of style, affect, and imagery in contemporary poetry. *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature* (pp. 8–17).

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 8788–8790.

Lee, C. M., Kilmer, J. R., Neighbors, C., Atkins, D. C., Zheng, C., Walker, D. D., & Larimer, M. E. (2013). Indicated prevention for college student marijuana use: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *81*, 702–709.

Lee, C. M., Neighbors, C., Lewis, M. A., Kaysen, D., Mittmann, A., Geisner, I. M., ... Larimer, M. E. (2014). Randomized controlled trial of a Spring Break intervention to reduce high-risk drinking. *Journal of Consulting and Clinical Psychology*, *82*, 189–201.

Lord, S. P., Can, D., Yi, M., Marin, R., Dunn, C. W., Imel, Z. E., ... Atkins, D. C. (2014). Advancing methods for reliably assessing motivational interviewing fidelity using the motivational interviewing skills code. *Journal of Substance Abuse Treatment*, *49*, 50–57.

Lundahl, B. W., Kunz, C., Brownell, C., Tollefson, D., & Burke, B. L. (2010). A meta-analysis of motivational interviewing: Twenty-five years of empirical studies. *Research on Social Work Practice*, *20*, 137–160.

Magill, M., Gaume, J., Apodaca, T. R., Walthers, J., Mastroleo, N. R., Borsari, B., & Longabaugh, R. (2014). The technical hypothesis of motivational interviewing: A meta-analysis of MI's key causal model. *Journal of Consulting and Clinical Psychology*, *82*, 973–983.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the ACL: System Demonstrations* (pp. 55–60).

Maskit, B., Bucci, W., & Murphy, S. (2015). A computer program for tracking the evolution of a psychotherapy treatment. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology, 2015* (pp. 134–145).

Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)* (pp. 1–12).

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of NAACL-HLT, 2013* (pp. 746–751).

Miller, W. R., Moyers, T. B., Ernst, D., & Amrhein, P. (2008). Manual for the Motivational Interviewing Skill Code (MISC), Version 2.0. Albuquerque, NM: Center on Alcoholism. *Substance Abuse, and Addictions, The University of New Mexico*.

Miller, W. R., & Rollnick, S. (2012). *Motivational interviewing: Helping people change* (3rd ed.). New York: Guilford Press.

Miller, W. R., & Rose, G. (2009). Toward a theory of motivational interviewing. *American Psychologist*, *64*, 527–537.

Miller, W. R., Sorensen, J. L., Selzer, J. A., & Brigham, G. S. (2006). Disseminating evidence-based practices in substance abuse treatment: A review with suggestions. *Journal of Substance Abuse Treatment*, *31*, 25–39.

Moyers, T., Martin, T., Catley, D., Harris, K. J., & Ahluwalia, J. S. (2003). Assessing the integrity of motivational interviewing interventions: Reliability of the motivational interviewing skills code. *Behavioural and Cognitive Psychotherapy*, *31*, 177–184.

Moyers, T. B., Martin, T., Houck, J. M., Christopher, P. J., & Tonigan, J. S. (2009). From in-session behaviors to drinking outcomes: A causal chain for motivational interviewing. *Journal of Consulting and Clinical Psychology*, *77*, 1113.

Neighbors, C., Lee, C. M., Atkins, D. C., Lewis, M. A., Kaysen, D., Mittmann, A., ... Larimer, M. E. (2012). A randomized controlled trial of event-specific prevention strategies for reducing problematic drinking associated with 21st birthday celebrations. *Journal of Consulting and Clinical Psychology*, *80*, 850–862.

Newman, D., Chemudugunta, C., Smyth, P., & Steyvers, M. (2006). Analyzing entities and topics in news articles using statistical topic models. *IEEE International Conference on Intelligence and Security Informatics*.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of ACL* (pp. 115–124).

Pennington, J., Socher, R., & Manning, C. D. (2014). *GloVe: Global Vectors for Word Representation.* In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing.

Rouhizadeh, M., Prud'hommeaux, E., Roark, B., & van S., J. (2013). Distributional semantic models for the evaluation of disordered speech. *Proceedings NAACL-HLT 2013* (pp. 709–714).

Roy-Byrne, P., Bumgardner, K., Krupski, A., Dunn, C., Ries, R., Donovan, D., ... Zarkin, G. A. (2014). Brief intervention for problem drug use in safety-net primary care settings. *JAMA*, *312*, 492–501.

Schwalbe, C. S., Oh, H. Y., & Zweben, A. (2014). Sustaining motivational interviewing: A meta-analysis of training studies. *Addiction*, *109*, 1287–1294.

Schwartz, A. H., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., ... Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One*, *8*, e73791.

Socher, R. (2014). *Recursive Deep Learning for Natural Language Processing and Computer Vision (Doctoral dissertation)*. Stanford: Palo Alto, CA.

Sridhar, V. K. R., Bangalore, S., & Narayanan, S. S. (2008). Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. *IEEE Transactions on Audio, Speech and Language Processing*, *16*, 797–811.

Substance Abuse and Mental Health Services Administration (2014). *National survey of substance abuse treatment services (N-SSATS): 2013 data on substance abuse treatment facilities.* Rockville, MD: author.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*, 24–54.

Tollison, S. J., Lee, C. M., Neighbors, C., Neil, T. A., Olson, N. D., & Larimer, M. E. (2008). Questions and reflections: The use of motivational interviewing microskills in a peer-led brief alcohol intervention for college students. *Behavior Therapy*, *39*, 183–194.

White, B., Lepreau, J., Stoller, L., Ricci, R., Guruprasad, S., Newbold, M., & Joglekar, A. (2002). An integrated experimental environment for distributed systems and networks. *OSDI02* (pp. 255–270) (Boston, MA).

Xiao, B., Bone, D., Van Segbroeck, Imel, Z. E., Atkins, D. A., Georgiou, P., & Narayanan, S. (2014). Modeling therapist empathy through prosody in drug addiction counseling. *Proceedings of Interspeech, 2014* (pp. 213–217).

Xiao, B., Georgiou, P., Imel, Z. E., Atkins, D., & Narayanan, S. S. (2013). Modeling therapist empathy and vocal entrainment in drug addiction counseling. *Proceedings of Interspeech, 2013* (pp. 2861–2865).

Xiao, B., Imel, Z. E., Atkins, D., Georgiou, P., & Narayanan, S. S. (2015). Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling. *Proceedings of Interspeech. Dresden, Germany*.