Building a Motivational Interviewing Dataset

Verónica Pérez-Rosas[1], Rada Mihalcea[1], Kenneth Resnicow[2]
Satinder Singh[1], Lawrence An[3]
[1]Computer Science and Engineering, University of Michigan
[2]School of Public Health, University of Michigan
[3]Center for Health Communications Research, University of Michigan
{vrncapr,mihalcea,kresnic,baveja,lcan}@umich.edu

Abstract

This paper contributes a novel psychological dataset consisting of counselors' behaviors during Motivational Interviewing encounters. Annotations were conducted using the Motivational Interviewing Integrity Treatment (MITI). We describe relevant aspects associated with the construction of a dataset that relies on behavioral coding such as data acquisition, transcription, expert data annotations, and reliability assessments. The dataset contains a total of 22,719 counselor utterances extracted from 277 motivational interviewing sessions that are annotated with 10 counselor behavioral codes. The reliability analysis showed that annotators achieved excellent agreement at session level, with Intraclass Correlation Coefficient (ICC) scores in the range of 0.75 to 1, and fair to good agreement at utterance level, with Cohen's Kappa scores ranging from 0.31 to 0.64.

Behavioral interventions are a promising approach to address public health issues such as smoking cessation, increasing physical activity, and reducing substance abuse, among others (Resnicow et al., 2002). In particular, Motivational Interviewing (MI), a client centered psychotherapy style, has been receiving increasing attention from the clinical psychology community due to its established efficacy for treating addiction and other behaviors (Moyers et al., 2009; Apodaca et al., 2014; Barnett et al., 2014; Catley et al., 2012).

Despite its potential benefits in combating addiction and in providing broader disease prevention and management, implementing MI counseling at larger scale or in other domains is limited by the need for human-based evaluations. Currently, this requires a human either watching or listening to video-tapes and then providing evaluative feedback.

Recently, computational approaches have been proposed to aid the MI evaluation process (Atkins et al., 2014; Xiao et al., 2014; Klonek et al., 2015). However, learning resources for this task are not readily available. Having such resources will enable the application of data-driven strategies for the automatic coding of counseling behaviors, thus providing researchers with automatic means for the evaluation of MI. Moreover, this can also be useful to explore how MI works by relating MI behaviors to health outcomes, and to provide counselors with evaluative feedback that helps them improve their MI skills.

In this paper, we present the construction and validation of a dataset annotated with counselor verbal behaviours using the Motivational Interviewing Treatment Integrity 4.0 (MITI), which is the current gold standard for MI-based psychology interventions. The dataset is derived from 277 MI sessions containing a total of 22,719 coded utterances.

1 Motivational Interviewing

Miller and Rollnick define MI as a collaborative, goal-oriented style of psychotherapy with particular attention to the language of change (Miller and Rollnick, 2013). MI has been widely used as a treatment method in clinical trials on psychotherapy research to address addictive behaviors such as alcohol, tobacco and drug use; promote healthier habits such as nutrition and fitness; and help clients with

Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 42–51,
San Diego, California, June 16, 2016. ©2016 Association for Computational Linguistics

psychological problems such as depression and anxiety disorders (Rollnick et al., 2008; Lundahl et al., 2010). In addition, MI has been successfully applied in different practice settings including social work in behavioral health centers, education, and criminal justice (Wahab, 2005; McMurran, 2009).

The competence of the counselor in MI delivery is measured using systematic observational methods to assess verbal behavior in MI by either focusing on therapist behaviors, client behaviors, or both (Jelsma et al., 2015). Current coding instruments for MI include the Behavior Change Counselor Index (BECCI) (Lane et al., 2005), the Client Evaluation of Motivational Interview (CEMI) (Madson et al., 2009), the Independent Tape Rating Scale (ITRS) (Martino et al., 2009), the MI Skills Code (MISC) (Moyers et al., 2003), the Stimulated Client Interview Rating Scale (SCIRS) (Arthur, 1999), the One Pass (McMaster and Resnicow, 2015), and the Motivational Interviewing Treatment Integrity (MITI) (Moyers et al., 2005).

1.1 Motivational Interviewing Treatment Integrity

The MITI coding system is currently the most frequently used instrument for assessing MI fidelity (Moyers et al., 2003). The MITI is derived from the MISC coding system and focuses exclusively on the verbal behavior of the counselor. It measures how well or poorly the clinician is using MI. The coding system evaluates MI processes related to change talk such as engagement, focus, evocation, and planning. MITI has two components: global scores and behavior counts. The global scores aim to characterize the overall quality of the interaction and include four dimensions, namely Cultivating Change Talk, Softening Sustain Talk, Partnership, and Empathy. Behavior counts are evaluated by tallying instances of particular interviewing behaviors, which can be grouped into five broad categories: questions, reflections, MI adherent behavior (MIA), MI non-adherent behavior (MINA), and neutral behaviors.

Reflections capture reflective listening statements made by the clinician in response to client statements and can be categorized as simple or complex. MIA behaviors summarize counselor adherence to core aspects of the MI strategy such as seeking collaboration, affirming, and emphasizing autonomy. MINA includes aspects that indicate counselor deficiencies while delivering MI, such as confronting and persuading without permission. The neutral behaviors include counselor actions such as providing information and persuading with permission.

MITI evaluation is conducted by trained coders who assess the overall session scores and the occurrence of behaviors by using pen and paper. During the coding process, coders rely on audio recordings and their corresponding transcriptions. The evaluation is usually performed as a two-step process by first evaluating overall scores and next focusing on behavior counts.

MITI coding is a very time consuming and expensive process, as it requires accurate transcriptions and human expertise. The quality of the transcriptions is affected by the recoding quality and their preparation is time consuming as it might take about three times the duration of the recording (Klonek et al., 2015). Thus, estimates for a 30 min session might add up to 2.5 hours of transcriber time and about one hour of coder time.

1.2 MI reliability assessment

Reliability assessment for MI helps to validate treatment fidelity in clinical studies as it provides evidence that the MI intervention has been effective and allows comparisons across studies (Jelsma et al., 2015). MI literature suggests assessing reliability by double coding a fraction of the study sessions. The most common method to quantify the inter-annotator agreement on MI coding is computing the Intraclass Correlation Coefficient (ICC). This statistic describes how much of the total variation in MITI scores is due to differences among annotators (Dunn et al., 2015). ICC scores range in the 0 to 1 interval; relatively high ICC scores indicate that annotators scored MITI in a similar way while scores closer to 0 suggest that there is a considerable amount of variation in the way annotator's evaluated counselor MI skill. Low scores further suggest that either the measure is defective or the annotators should be retrained. Another method to measure inter-annotator reliability in MI is the Cohen's Kappa score (Lord et al., 2015a), which calculates the pair-wise agreement among annotations considering the probability of annotators agreeing by chance.

2 Related work

Current approaches for MI coding and evaluation entail extensive human involvement. Recently, there have been a number of efforts on building computational tools that assist researchers during the coding process. (Can et al., 2012) proposed a linguistic based approach to automatically detect and code counselor reflections that is based on analyzing n-grams patterns, similarity features between counselor and client speech, and contextual meta-features, which aim to represent the dialog sequence between the client and counselor. A method based on topic models is presented in (Atkins et al., 2012; Atkins et al., 2014), where authors focus on automatically identifying topics related to MI behaviors such as reflections, questions, support, and empathy, among others. Text and speech based methods have also been proposed to evaluate overall MI quality. (Lord et al., 2015b) analyzed the language style synchrony between therapist and client during MI encounters. In this work, authors relied in the psycholinguistic categories from the Linguistic Inquiry and Word Count lexicon to measure the degree in which counselor matches the client language. (Xiao et al., 2014) presents a study on the automatic evaluation of counselor empathy by analyzing correlations between prosody patterns and empathy showed by the therapist during the counseling interaction.

Although most of the work on coding of MI within session language has focused on modeling the counselor language, there is also work that addresses the client language. (Tanana et al., 2015) used recursive neural networks (RNN) to identify client change and sustain talk in MI transcripts, i.e., language that indicates commitment towards and away behavioral change. In this work, authors combined both therapist and client utterances in a single sequence model using Maximum Entropy Markov Models, NRR, and n-grams features. (Gupta et al., 2014) analyzed the valence of client's attitude towards the target behavior by using n-grams and conditional maximum entropy models. In this paper authors also present an exploration of the role laughter of both counselor and client's during the MI encounter and attempts to incorporate its occurrence as an additional source of information in the prediction model.

Research findings have shown that natural language processing approaches can be successfully applied to behavioral data for the automatic annotation of therapists' and clients' behaviors. This motivates our interest in building resources for this task as an initial step for the construction of improved coding tools. There has been work on creating annotated resources that facilitate advances in natural language processing of clinical text, including semantic and syntactic annotation of pathology reports, oncology reports, and biomedical journals (Roberts et al., 2007; Albright et al., 2013; Verspoor et al., 2012). However, to our knowledge, there are just a few psychotherapy corpora available. One of them is the "Alexander Street Press", [1] which is a large collection of transcripts and video recordings of therapy sessions on different subjects such as anxiety, depression, family conflicts, and others. There are also some other psychology datasets available under limited access from the National Institute of Mental Health (NIMH).[2] These datasets provide recorded interactions among clinicians and patients on a number of psychology styles. However, they are not annotated and validated to be used in the computational psychology domain.

In this paper, we present the development of a clinical narratives dataset that can be used to implement data-driven methods for the automatic evaluation of MI sessions.

3 Motivational Interviewing Dataset

3.1 Data collection

The dataset is derived from a collection of 284 video recordings of counseling encounters using MI. The recordings were collected from various sources, including two clinical trials, students' counseling sessions from a graduate level MI course, wellness coaching phone calls, and demonstrations of MI strategies in brief medical encounters.

The clinical trials sessions consist of interventions for smoking cessation and antiretroviral therapy adherence with electronic drug monitoring. Psychology students' sessions are conducted on standardized patients and aim at weight loss and smoke

[1] http://alexanderstreet.com/products/counseling-and-psychotherapy-transcripts-series

[2] http://psychiatry.yale.edu/pdc/resources/datasets.aspx

Source	No. sessions	Avg.length
Clinical trial	121	27 min
Standardized patients	138	15 min
Brief MI encounters	18	4 min
Coaching phone calls	7	15 min
Total	284	

Table 1: Data sources for the MI sessions

cessation. Wellness coaching phone calls inquired about patient health and medication adherence after surgery. The demonstration sessions are collected from online sources, i.e., YouTube and Vimeo, and present brief MI encounters on several scenarios such as dental practice, emergency room counseling, and student counseling. Table 1 presents a summary of the data sources used in the dataset collection.

All the sessions are manually anonymized to remove identifiable information such as counselor and patient names and references to counseling sites' location. Each recording is assigned a new identifier that does not include any information related to the original recording. The resulting recordings are then processed to remove the visual data stream to further prevent counselor/patient identification. After this process, we obtained a set of 277 sessions due to the exclusion of some sessions with recording errors. The final dataset comprises a total of 97.8 hours of audio with average session duration of 20.8 minutes and a standard deviation of 11.4 minutes.

3.2 Transcriptions

The sessions from clinical trials include full transcripts. However, this was not the case for the remaining set of sessions, and for these we obtained manual transcriptions via crowdsourcing using Amazon Mechanical Turk. This resource has proved to be a fast and reliable method to obtain speech transcriptions (Marge et al., 2010).

Mechanical Turk workers were provided with transcription guidelines that include clearly identifying the speaker (client or counselor), and transcribing speech disfluencies such as false starts, repetitions of whole words or parts of words, prolongations of sounds, fillers, and long pauses. Resulting transcriptions were manually verified to avoid spam and to ensure their quality. The transcriptions consist of approximately 22,719 utterances, with an average of 83 utterances per session.

3.3 MITI Annotations

Three counselors, with previous MI experience, were trained on the use of the MITI 4.1 by expert trainers from the Motivational Interviewing Network of Trainers[3] (MINT) to conduct the annotation task. Prior to the annotation phase, annotators participated in a coding calibration phase where they had discussions regarding the criteria for sentence parsing, the correct assignment of behavior codes, and conducted team coding of sample sessions.

Annotators used both audio recordings and verbatim transcriptions to conduct the annotations.

Annotators were instructed to parse the interviewer speech following the guidelines defined by MITI 4.1. The annotation was conducted at utterance level, by selecting and labeling utterances in each counselor turn that contain a specific MI behavior.

Following this strategy allowed us to obtain more accurate examples of each behavior code for cases where a turn contains multiple utterances and thus more than one behavior code. In addition, given possible inaccuracies and interruptions in the turn by turn segmentation, annotators were allowed to select the text that they considered belonging to a coded utterance, even if it spanned more than one counselor-client turn, to avoid utterance breaking.

In order to facilitate this process, annotators used a software based coding system instead of the traditional paper and pen system. Annotators were trained to code using the Nvivo software,[4] a quantitative analysis suite that allows to select and assign text segments to a given codebook. The codebook contains the following behavior codes:

Question (QUEST) All questioning statements spoken by clinicians.

Simple reflection (SR) Clinician statements that convey understanding or facilitate client-clinician exchanges.

Complex reflection (CR) Reflective statements that add substantial meaning or emphasis to what the client has said.

[3] http://www.motivationalinterviewing.org/

[4] http://www.qsrinternational.com/what-is-nvivo

Code	Count	Verbal example
QUEST	5262	Could you talk a little bit more about those behaviors you say that automatically makes you smoke?
SR	2690	It sounds like something that you know and feel like you can improve on in the next week.
CR	2876	So you want something that's gonna to allow you to eat the foods that you enjoy but that maybe more moderation.
SEEK	614	And, then, when we meet again, you can bring some of that information. Maybe we can discuss which of those feels right for you, and start to put together a plan for what could be your next steps.
AUTO	141	This is something that it's up to you whether you want to use it or not.
AF	499	Okay, great. So, I'm excited about this because you're obviously very motivated. And the barriers that you've presented are definitely overcomable
CON	141	Okay, well that's a good start, but cutting back isn't gonna do it. If you actually quit the smoking, you can reverse all the damage you've done in your mouth, and you can stop yourself from ... from being at risk for these other diseases. But, but as long as you're continuing to use these cigars, you're really putting yourself in a lot of danger.
PWOP	598	Okay so with all of the risks of smoking and the benefits of quitting, what is keeping you from making a plan?
N-GI	1017	There are two other over the counter options. There's a patch and that would deal with the taste you don't like. With the patch you just put it on and it slowly releases nicotine throughout the day so you don't even have to think about it. There are also lozenges, which are kind of like throat lozenges, or a hard candy and you just suck on it. And as it dissolves it releases nicotine.
N-PWP	2100	Well, if it's alright with you, umm, you know, I could toss out some ideas of things that have worked for other people and umm things that umm, could be helpful as far as reducing stress and, and really filling in other activities so you're not umm, as tempted to ... smoke

Table 2: Frequency counts and verbal examples of MI behaviors in the dataset

Seeking collaboration (SEEK) The clinician attempts to share power or acknowledge the expertise of the client.

Emphasizing autonomy (AUTO) The clinician focus the responsibility on the client for the decision and actions pertaining to change.

Affirm (AF) Clinician utterances that accentuates something positive about the client.

Persuading without permission (PWOP) The clinician attempts to change the client's opinions, attitudes, or behaviors, using tools such as logic, compelling arguments, self-disclosure or facts.

Confront (CON) Statements where the clinician confronts the client by directly disagreeing, arguing, correcting, shaming, criticizing, moralizing or questioning client's honesty.

Persuading with permission (N-PWP) Clinician statements that make emphasis on collaboration or autonomy support while persuading.

Giving information (N-GI) The clinician give information, educates, or expresses a professional opinion without persuading, advising, or warning.

The 277 sessions were randomly distributed among the three annotators. The team annotated approximately 20 sessions per week. The entire annotation process took about three months.

Annotator 1		Annotator 2		Method
So you're getting back to your old self.	SR	So you're getting back to your old self.	SR	Exact match
So it sounds like you kinda struggle with that a little bit in that sometimes	SR	So it sounds like you kinda struggle with that a little bit in that sometimes it's hard I imagine, it is sometimes hard to be financially independent I mean I, But it something it sounds like you respect in yourself that you are able to do it.	SR	Split utterances
it's hard I imagine, it is sometimes hard to be financially independent I mean I,	NL			
But it something it sounds like you respect in yourself that you are able to do it.	SR			
OK. But even though it's something that you really don't like, it's something that's not terribly bothersome.	SR	So you mentioned that one side effect of the Sustiva was that it makes you dizzy. OK. But even though it is something that you really don't like, it something that,it's not terribly bothersome.	SR	Partial match

Table 3: Sample utterance alignment for coding comparisons

After the annotation phase, transcripts were processed to extract the verbal content of each MITI annotation; non-coded utterances were also extracted and labeled as neutral. Sample annotations are presented in Table 2. The final set contains 15,886 annotations distributed among the ten codes and 6,833 neutral utterances. Table 2 shows the frequency distribution for each behavior count and neutral utterances.

4 Dataset Validation

In order to validate the annotator reliability, a sample of 10 sessions was randomly selected to be double coded by two members of the coding team.

The total amount of recoding material for this sample is about 4.5 hours. Each session has an average duration of 26 minutes, with an average of 115 counselor-client conversation turns per session, comprising a total of 1,160 utterances.

4.1 Inter-rater Reliability Analysis

Because we conducted the MITI annotation at utterance level without any pre-parsing, annotations across coders showed noticeable parsing variations. These variations consisted of differences in utterance boundaries such as overlaps and split utterances. In order to allow for coding comparisons, we opted for aligning annotations by utterance matching using similar methods to (Lord et al., 2015a). We considered three cases: exact match, partial match and split utterances. In the first case, we simply compare two coded utterances and define a match if both utterances contained the same words. The partial match addresses cases where two coders dis-

Code	ICC	Kappa
QUEST	0.97	0.64
CR	0.97	0.49
SR	0.89	0.34
SEEK	0.03	0.42
N-GI	0	0.28
AF	0	0.47
AUTO	0	0.31
N-PWP	0	NA
CON	NA	NA
PWOP	NA	NA

Table 4: ICC at session level and Kappa scores at utterance level for 10 double coded sessions. NA indicates that the MI behavior was not present in any session

agree in utterance boundaries, thus resulting in annotations from one annotator partially matching the others, i.e., some degree of overlap. The third case also deals with differences due to utterance boundaries but focuses on split utterances, i.e., an annotated utterance from one coder was split into two different annotations by the other, and cases where utterances with different annotations show some degree of overlapping. Table 3 presents sample utterances.

Using the transcript from each session, we first identified those utterances who were assigned a behavior code by either of the two annotators. Then, we compared their verbal content by applying the utterance matching methods described above. We assigned a match when both annotators agreed on their evaluations. We considered both split utterances and partial matches as a single match. Those utterances for which we were unable to find a matching pair or differed on the assigned codes were regarded as disagreements.

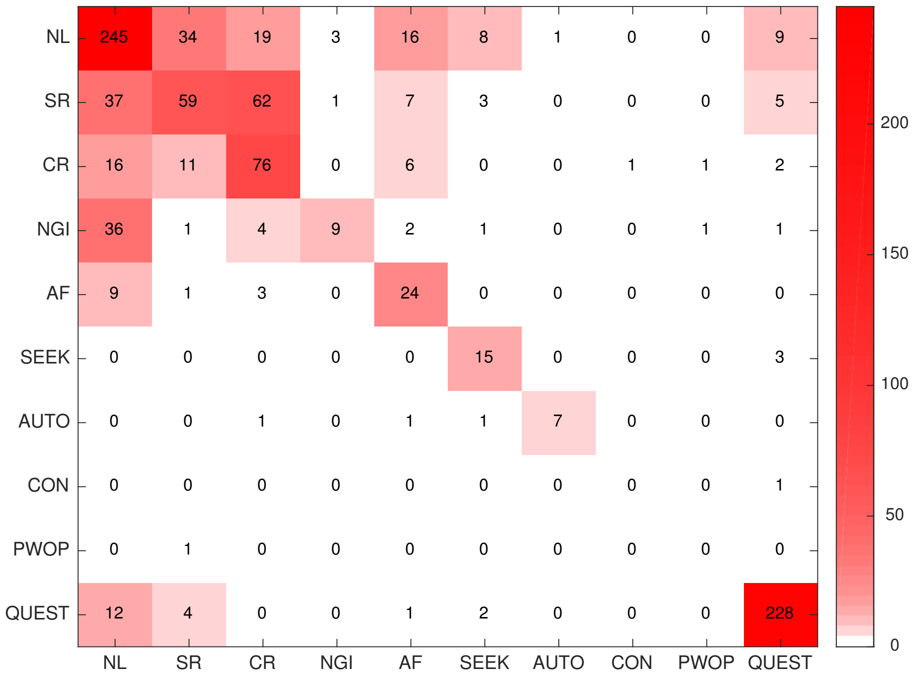
Table 4 presents the Intra Class Coefficient (ICC) measured at session level. Reported scores were obtained using a two-way mixed model with absolute agreement (Jelsma et al., 2015). Overall, we observe excellent ICC scores for Complex Reflections CR (CR), Simple Reflections (SR), and Questions, based on ICC reference values, where values ranging from 0.75 to 1 are considered as excellent agreement (Jelsma et al., 2015).

ICC scores suggest that annotators did not show significant variations on most of the coded behaviours, except for Seeking collaboration (SEEK), which showed considerable disagreement. We believe that this is caused by the higher variability on the frequency counts for this code across the 10 sessions.

Wanting to evaluate how well did the annotators agree while coding the same annotation, we calculated the pairwise agreement among coders using Cohen's Kappa. Results are also reported in Table 4. The Kappa values suggest fair to good levels of agreement among the different behavior codes.

In addition, we evaluate the ability of coders to distinguish the occurrence of a particular behavior code versus any other code. This allow us to answer question such as, how well did the annotators agree on what is considered a reflection as compared to what is not a reflection? This analysis provides further insights about the validity of the coding. In these comparisons, utterances coded with a different behavior than the target behavior were considered as the negative case. For instance, if the target behavior was Simple Reflection (SR), then we evaluated the identification of Simple Reflection vs non-Simple Reflection. In order to more accurately represent the human coding process, we also included non coded utterances (NL) as negatives cases. Figure 1 shows the annotation agreement between the two annotators for 10 sessions coded at utterance level in heatmap representation, where the color intensity represents the agreement distribution. In the shown matrix, the *x* axis indicates the MI code assigned by the first annotator and the *y* axis the MI code assigned by the second annotator. Each cell indicates the observed frequency of a coding pair.

From this table, we observe that questions attained the highest agreement levels among all behaviors, followed by simple reflections (SR), complex reflections (CR), seeking collaboration (SEEK), giving information (GI), and emphasizing autonomy (AUTO). From the observed disagreements, a small fraction of questions annotated by one coder were regarded as Simple Reflections or were left uncoded by the second coder. This might be related to ambiguous cases, where the counselor formulate a simple reflection but added a question tone at end of the sentence thus making the reflection sound like a question. In addition, annotators showed noticeable disagreement while distinguishing between complex and simple reflections. This was somehow expected, as the MI literature has reported similar findings given the highly subjective criteria applied while evaluating these codes (Lundahl et al., 2010). Annotators found no agreement for confronting (CON) and persuading without permission (PWOP) codes. This has to do with zero or low frequency counts e.g the first annotator found only one confrontation utterance while the second annotator found zero. Finally, annotators showed high agreement on utterances that did not contain MI behaviors, thus suggesting that 1) annotators have good agreement regarding what should be coded; and 2) differences in parsing did not affect the annotations process.



	NL	SR	CR	NGI	AF	SEEK	AUTO	CON	PWOP	QUEST
NL	245	34	19	3	16	8	1	0	0	9
SR	37	59	62	1	7	3	0	0	0	5
CR	16	11	76	0	6	0	0	1	1	2
NGI	36	1	4	9	2	1	0	0	1	1
AF	9	1	3	0	24	0	0	0	0	0
SEEK	0	0	0	0	0	15	0	0	0	3
AUTO	0	0	1	0	1	1	7	0	0	0
CON	0	0	0	0	0	0	0	0	0	1
PWOP	0	1	0	0	0	0	0	0	0	0
QUEST	12	4	0	0	1	2	0	0	0	228

Figure 1: Annotator agreement on non-coded utterances (NL) and MI behaviors. The *x* axis indicates the MI code assigned by the first annotator and the *y* axis the MI code assigned by the second annotator.

5 Conclusion

In this paper, we introduced a new clinical narratives dataset derived from MI interventions. The dataset consists of annotations for ten verbal behaviors displayed by the counselor while conducting MI counseling. We presented a detailed description of the dataset collection and annotation process. We conducted a reliability analysis where we showed that annotators achieved excellent agreement at session level, with ICC scores in the range of 0.75 to 1, and fair to good agreement at utterance level, with Cohen's Kappa scores ranging from 0.31 to 0.64. The paper reports our initial efforts towards building accurate tools for the automatic coding of MI encounters. Our future work includes developing data-driven methods for the prediction of MI behaviors.

Acknowledgments

This material is based in part upon work supported by the University of Michigan under the MCubed program, by the National Science Foundation award #1344257, and by grant #48503 from the John Templeton Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the University of Michigan, the National Science Foundation, or the John Templeton Foundation.

References

Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F Styler, Colin Warner, Jena D Hwang, Jinho D Choi, Dmitriy Dligach, Rodney D Nielsen, James Martin, et al. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative.

Timothy R Apodaca, Brian Borsari, Kristina M Jackson, Molly Magill, Richard Longabaugh, Nadine R Mastroleo, and Nancy P Barnett. 2014. Sustain talk predicts poorer outcomes among mandated college student drinkers receiving a brief motivational intervention. *Psychology of Addictive Behaviors*, 28(3):631.

David Arthur. 1999. Assessing nursing students' basic communication and interviewing skills: the development and testing of a rating scale. *Journal of Advanced Nursing*, 29(3):658–665.

David C Atkins, Timothy N Rubin, Mark Steyvers, Michelle A Doeden, Brian R Baucom, and Andrew Christensen. 2012. Topic models: A novel method for modeling couple and family text data. *Journal of family psychology*, 26(5):816.

David C Atkins, Mark Steyvers, Zac E Imel, and Padhraic Smyth. 2014. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, 9(1):49.

Elizabeth Barnett, Theresa B Moyers, Steve Sussman, Caitlin Smith, Louise A Rohrbach, Ping Sun, and Donna Spruijt-Metz. 2014. From counselor skill to decreased marijuana use: Does change talk matter? *Journal of substance abuse treatment*, 46(4):498–505.

Dogan Can, Panayiotis G. Georgiou, David C. Atkins, and Shrikanth S. Narayanan. 2012. A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features. In *INTERSPEECH*, pages 2254–2257. ISCA.

Delwyn Catley, Kari J Harris, Kathy Goggin, Kimber Richter, Karen Williams, Christi Patten, Ken Resnicow, Edward Ellerbeck, Andrea Bradley-Ewing, Domonique Malomo, et al. 2012. Motivational interviewing for encouraging quit attempts among unmotivated smokers: study protocol of a randomized, controlled, efficacy trial. *BMC public health*, 12(1):456.

Chris Dunn, Doyanne Darnell, Sheng Kung Michael Yi, Mark Steyvers, Kristin Bumgardner, Sarah Peregrine Lord, Zac Imel, and David C Atkins. 2015. Should we trust our judgments about the proficiency of motivational interviewing counselors? a glimpse at the impact of low inter-rater reliability. *Motivational Interviewing: Training, Research, Implementation, Practice*, 1(3):38–41.

R Gupta, P G Georgiou, D C Atkins, and S Narayanan. 2014. Predicting client's inclination towards target behavior change in motivational interviewing and investigating the role of laughter. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (September):208–212.

Judith GM Jelsma, Vera-Christina Mertens, Lisa Forsberg, and Lars Forsberg. 2015. How to measure motivational interviewing fidelity in randomized controlled trials: Practical recommendations. *Contemporary clinical trials*, 43:93–99.

Florian E Klonek, Vicenç Quera, and Simone Kauffeld. 2015. Coding interactions in motivational interviewing with computer-software: What are the advantages for process researchers? *Computers in Human Behavior*, 44:284–292.

Claire Lane, Michelle Huws-Thomas, Kerenza Hood, Stephen Rollnick, Karen Edwards, and Michael Robling. 2005. Measuring adaptations of motivational interviewing: the development and validation of the behavior change counseling index (becci). *Patient education and counseling*, 56(2):166–173.

Sarah Peregrine Lord, Doğan Can, Michael Yi, Rebeca Marin, Christopher W Dunn, Zac E Imel, Panayiotis Georgiou, Shrikanth Narayanan, Mark Steyvers, and David C Atkins. 2015a. Advancing methods for reliably assessing motivational interviewing fidelity using the motivational interviewing skills code. *Journal of substance abuse treatment*, 49:50–57.

Sarah Peregrine Lord, Elisa Sheng, Zac E Imel, John Baer, and David C Atkins. 2015b. More than reflections: empathy in motivational interviewing includes language style synchrony between therapist and client. *Behavior therapy*, 46(3):296–303.

Brad W Lundahl, Chelsea Kunz, Cynthia Brownell, Derrik Tollefson, and Brian L Burke. 2010. A meta-analysis of motivational interviewing: Twenty-five years of empirical studies. *Research on Social Work Practice*.

Michael B Madson, E Bullock, A Speed, and S Hodges. 2009. Development of the client evaluation of motivational interviewing. *Motivational Interveiwing Network of Trainers Bulletin*, 15:6–8.

Matthew Marge, Satanjeev Banerjee, Alexander Rudnicky, et al. 2010. Using the amazon mechanical turk for transcription of spoken language. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5270–5273. IEEE.

Steve Martino, Samuel A Ball, Charla Nich, Tami L Frankforter, and Kathleen M Carroll. 2009. Informal discussions in substance abuse treatment sessions. *Journal of substance abuse treatment*, 36(4):366–375.

Fiona McMaster and Ken Resnicow. 2015. Validation of the one pass measure for motivational interviewing competence. *Patient education and counseling*, 98(4):499–505.

Mary McMurran. 2009. Motivational interviewing with offenders: A systematic review. *Legal and Criminological Psychology*, 14(1):83–100.

William R Miller and Stephen Rollnick. 2013. *Motivational interviewing: Helping people change, Third edition*. The Guilford Press.

Theresa Moyers, Tim Martin, Delwyn Catley, Kari Jo Harris, and Jasjit S Ahluwalia. 2003. Assessing the integrity of motivational interviewing interventions: Reliability of the motivational interviewing skills code. *Behavioural and Cognitive Psychotherapy*, 31(02):177–184.

Theresa B Moyers, Tim Martin, Jennifer K Manuel, Stacey ML Hendrickson, and William R Miller. 2005.

Assessing competence in the use of motivational interviewing. *Journal of substance abuse treatment*, 28(1):19–26.

Theresa B Moyers, Tim Martin, Jon M Houck, Paulette J Christopher, and J Scott Tonigan. 2009. From in-session behaviors to drinking outcomes: a causal chain for motivational interviewing. *Journal of consulting and clinical psychology*, 77(6):1113.

Ken Resnicow, Colleen DiIorio, Johana E Soet, Belinda Borrelli, Denise Ernst, Jacki Hecht, and Angelica Thevos. 2002. Motivational interviewing in medical and public health settings. *Motivational interviewing: Preparing people for change*, 2:251–269.

Angus Roberts, Robert Gaizauskas, Mark Hepple, Neil Davis, George Demetriou, Yikun Guo, Jay Kola, Ian Roberts, Andrea Setzer, Archana Tapuria, and Bill Wheeldin. 2007. The CLEF corpus: semantic annotation of clinical text. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 625–629.

Stephen Rollnick, William R Miller, Christopher C Butler, and Mark S Aloia. 2008. Motivational interviewing in health care: helping patients change behavior. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 5(3):203–203.

Michael Tanana, Kevin Hallgren, Zac Imel, David Atkins, Padhraic Smyth, and Vivek Srikumar. 2015. Recursive Neural Networks for Coding Therapist and Patient Behavior in Motivational Interviewing. *2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 71–79.

Karin Verspoor, Kevin Bretonnel Cohen, Arrick Lanfranchi, Colin Warner, Helen L Johnson, Christophe Roeder, Jinho D Choi, Christopher Funk, Yuriy Malenkiy, Miriam Eckert, et al. 2012. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC bioinformatics*, 13(1):1.

StÉphanie Wahab. 2005. Motivational interviewing and social work practice. *Journal of Social Work*, 5(1):45–60.

Bo Xiao, Daniel Bone, Maarten Van Segbroeck, Zac E Imel, David C Atkins, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2014. Modeling therapist empathy through prosody in drug addiction counseling. In *Fifteenth Annual Conference of the International Speech Communication Association*.