A Dialog Act Tagging Approach to Behavioral Coding: A Case Study of Addiction Counseling Conversations

Doğan Can[1], *David C. Atkins*[3] *and Shrikanth S. Narayanan*[1,2]

[1]Department of Computer Science and [2]Department of Electrical Engineering,
University of Southern California, Los Angeles, CA, USA
[3]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA

dogancan@usc.edu, datkins@u.washington.edu, shri@sipi.usc.edu

Abstract

Motivational Interviewing (MI) is a goal-oriented psychotherapy, employed in cases such as addiction, that helps clients explore and resolve their ambivalence about the problem at hand in a dialog setting. MI session quality is typically assessed with behavioral coding – a time consuming and labor intensive manual annotation system. This paper examines a computational approach to modeling and assessing the quality of MI sessions. Specifically, we pose the utterance level behavioral coding task as a sequence tagging problem and use linear chain CRF models trained on coded session transcripts and Switchboard DAMSL dataset to predict utterance level behavioral codes as well as dialog acts. We then use those utterance level predictions to predict session level behavioral codes of clinical interest characterizing the quality and efficacy of psychotherapy. We experiment with different feature parameterizations and reduced code sets and present an analysis of how standard dialog acts relate to behavioral codes.

Index Terms: dialog act tagging, behavioral coding, motivational interviewing skills code

1. Introduction

Motivational Interviewing (MI) is a non-confrontational, directive counseling approach extensively used in treating alcohol and other drug-related problems [1, 2]. Many individuals struggling with addictions perceive both benefit (e.g., the “high”) and harm (e.g., missing work, problems with spouses) from their use and are ambivalent about changing their behavior. MI focuses on eliciting and enhancing the intrinsic motivation for change by exploring and resolving client ambivalence in a dyadic spoken dialog setting. MI is client-focused; it has an empathic focus and emphasizes the client’s right and responsibility to make changes related to their addictive behaviors. There is a strong evidence-base for MI [3], and a current focus in mental health policy research is how to support the effective implementation and dissemination of MI into community treatment centers. One significant challenge is how to ensure high quality treatment. Heretofore, the typical method for assessing quality, or proficiency, of MI is to use a behavioral coding system in which human raters learn and then utilize a system for annotating video or audio tapes. However, this approach does not scale up to real-world use [4]. Thus, an automated solution to coding – and hence assessment of MI quality – would greatly enhance the dissemination of high-quality MI into the community.

There are two major approaches to coding MI sessions, the Motivational Interviewing Skills Code (MISC) [2] and the Motivational Interviewing Treatment Integrity (MITI) Code [5]. Both methods evaluate the quality of MI from audio- and/or video-tapes of individual counseling sessions but they aim to accomplish different tasks at different levels of resolution. MISC is typically used for detailed psychotherapy process research investigating the critical elements and causal mechanisms within MI that correlate with efficacy. It is designed for manually annotating an interview between two individuals, the Counselor and the Client. This annotation is performed by trained coders who associate an appropriate behavioral code to each counselor and client utterance in addition to assigning session level global ratings that characterize the entire interaction. Utterance level codes are similar in function to dialog acts that are used to annotate high-level discourse structure of a dialog [6]. They encode local events in the conversation that are of clinical interest, such as the counselor asking open ended questions inviting the client to open up and talk about their thoughts and feelings or the client talking about making a change for the better. While utterance level MISC coding is an invaluable resource for MI process research, it is at the same time a phenomenally labor intensive annotation task. MITI is essentially a stripped down version of MISC with a narrower scope and much lower resource requirements. It aims to answer simpler questions like “How much is this session like MI?”. MITI coding involves listening to a MI session recording, counting instances of a small subset of MISC codes, i.e. no utterance level annotation, and assigning global ratings characterizing the entire session. Both MISC and MITI use session level code tallies along with global ratings to measure MI session quality.

In this paper, we develop a two-stage computational approach to predicting session level MI quality measures, i.e. global ratings, from session transcripts. We investigate the viability of two scenarios:

1. Can we learn to predict session level MI quality measures from a new session transcript if we already have a large corpus of session transcripts annotated with MISC and MITI codes? In the first scenario, we model the utterance level MISC coding procedure as a sequence tagging problem and train a linear chain CRF model [7] on coded session transcripts for predicting utterance level MISC codes. We then use the utterance level code predictions to predict session level MI quality measures.
2. Can we learn to predict session level MI quality measures from a new session transcript if we only have a large corpus of session transcripts annotated with MITI

Research supported by NSF, NIH and DoD.

codes by exploiting the similarities between MISC codes and dialog acts? In the second scenario, we train a linear chain CRF model on the Switchboard DAMSL conversations for predicting dialog act tags. We then use the dialog act predictions to predict session level MI quality measures.

We experiment with different feature parameterizations, simpler behavioral code and dialog act sets and present an analysis of how dialog acts relate to behavioral codes.

2. Coding of Motivational Interviews

The standard MISC coding procedure consists of several passes through an interview in which each pass is aimed at coding a particular aspect of the session like the counselor behavior, client behavior or the overall efficacy. Although there is a certain amount of workflow variation between different MI studies that employ MISC, here we will gloss over the details of the actual coding process and limit our discussion to the aspects of MISC relevant to this study. The MISC coding output consists of annotations at two different resolutions: utterance-level behavioral codes and session-level global ratings on a 7-point Likert scale that are intended to be a holistic evaluation of the entire session, one that cannot necessarily be separated into individual elements [2].

In this study we are interested in two global ratings of counselor behavior commonly used in MISC and MITI coding: Empathy and Spirit. Both of these ratings are intended to capture the coder's overall impression of the counselor's performance throughout the session. Empathy rating is intended to capture the extent to which the counselor understands and/or makes an effort to accurately understand the client's perspective. Counselors high on this scale show an active interest in making sure they understand what the client is saying. Spirit rating, on the other hand, is intended to capture the overall competence of the counselor in using motivational interviewing [2].

The MISC defines an utterance as a *complete thought* and it ends either when the speaker completes a thought and moves to another or when the speaker changes. MISC 2.1 [2] differentiates between 15 major categories of counselor behavior and 5 major categories of client behavior at the utterance level (summarized in Table 1). Client language is divided into three groups. Any language that moves in the direction of change is termed "change talk", and any language moving in the opposite direction is termed "sustain talk". All MI sessions are conducted with a Target Behavior Change (TBC) in mind. In general TBC is the problem area specified by the research protocol and is the focus of a therapy session. Accordingly MISC only codes the change/sustain talk relevant to TBC. All client categories (except FN) require a valence rating depending on whether they reflect inclination toward (change talk, positive valence) or away from (sustain talk, negative valence) the TBC. All client speech that is not TBC-relevant change/sustain talk is coded as FN, the default category for client language. Utterance level codes are intended to capture specific local behaviors within the dialog. While the global context might influence the coder decisions, in general codes are determined by considering the local dialog context.

The MITI coding procedure keeps tallies of a limited grouping of utterance level MISC codes. Giving information, simple/complex reflections and open/closed questions are kept as separate categories, MI adherent (advise with permission, affirm, emphasize control, support) and MI non-adherent (advise w/o permission, confront, direct) behaviors are grouped, and rest of the MISC codes are ignored.

Table 1: MISC Categories

Code	Category	Count
	Counselor	
ADP	Advise with permission	105
ADW	Advise w/o permission	598
AF	Affirm	1649
CO	Confront	187
DI	Direct	134
EC	Emphasize Control	133
FA	Facilitate	16296
FI	Filler	157
GI	Giving Information	15748
QUC	Closed Question	5276
QUO	Open Question	4562
RCP	Raise Concern with permission	4
RCW	Raise Concern w/o permission	42
REC	Complex Reflection	4703
RES	Simple Reflection	6354
RF	Reframe	19
ST	Structure	1223
SU	Support	642
WA	Warn	65
	Client	
C±	Commitment	111/21
FN	Follow/Neutral	47491
R±	Reason	3278/2828
O±	Other	1788/1638
TS±	Taking Steps	133/51

Reflections (RES, REC) and questions (QUO, QUC) are believed to be critical components of MI [5, 2]. Reflections tend to have an empathic tone and can serve to reflect back to the client both positive and negative outcomes of their addictions – thus, helping the client to resolve (or at least face) their ambivalence. Simple Reflections (RES) are those which add little or no meaning to what the client has said. Their primary function is to convey understanding. Repeating or rephrasing what the client has said is considered RES. Complex Reflections (REC) typically add substantial meaning or emphasis to what the client has said. They convey a deeper or richer picture of the client's statement and may contain significantly more or different content from what the client has actually said. The differences may be subtle or obvious. Analogies, metaphors, similes, exaggerations and summaries almost always fall under the REC category. Closed Questions (QUC) imply a short answer while Open Questions (QUO) allow a wide range of possible answers. Open questions might seek information, invite the client's perspective or encourage self-exploration and they are thought to be a critical counselor behavior for successful MI.

3. Data

This section briefly describes the data sets used in experiments. Note that we performed standard normalization operations on these data sets prior to experiments. We removed all punctation except apostrophes and underscores (used for keeping entities together), tokenized utterances by splitting on white space, and finally lowercased everything.

3.1. MI Dataset

MI data includes 148 sessions from five separate MI intervention studies [8] focusing on drug and alcohol abuse problems and 195 sessions from a multi-site MI training study [9]. Sessions last from 20 minutes to an hour depending on the study. Prior to coding, each session audio-tape was carefully transcribed (marking back-channels, disfluencies, interruptions, overlaps, etc.) and annotated with turn-level time alignments. All sessions were MISC coded by trained coders who also segmented turns into MISC utterances, i.e. complete thoughts. Some sessions were coded multiple times to establish inter- and intra-coder reliability.

We use a fixed 70% train, 30% test split in all experiments. The training subset includes 232 sessions (117K utterances, 1.2M words). The test subset includes 111 sessions (58K utterances, 536K words). Individual counts of MISC codes in the train subset are given in the last column of Table 1. There are 1.9K utterances in the train set that were not assigned any MISC codes.

3.2. Switchboard-DAMSL Dataset

The Switchboard-DAMSL corpus [6] consists of 1155 dialogs (219K utterances) from the Switchboard telephone conversations. All dialogs are tagged with discourse labels at the utterance level. Although the original labels assigned by coders constitute a set of approximately 220 unique labels, these were later clustered into 42 mutually exclusive dialog acts to obtain enough data per label for statistical modeling. We use these 42 tags as well as the simpler 7 tags described in [10] in our experiments. We also use the train test split defined in [6] (1115 train dialogs, 19 test dialogs) for training and evaluating our dialog act tagging system.

4. Method

In this paper we predict session level MI quality measures using a 2-stage approach. In the first stage we tag each utterance with a label, i.e. a behavioral code or a dialog act, using a linear chain CRF model and then use session level tallies of those labels to predict binarized (High, Low) session level Empathy and Spirit ratings using logistic regression. We normalize the tallies for each session by dividing them with the total number of utterances in a session.

Linear chain CRFs [7] are used extensively in sequence tagging tasks due to the relative ease of incorporating arbitrary features of observations into prediction. We process utterances in sequence and extract binary indicator features for each utterance by combining observations from a parametrized local context with the current label (unigram) or the label bigram constructed by concatenating the labels for the current utterance and the previous one. At training time, these features combine reference labels with static observations. At decoding time, the same features combine hypothesized labels with static observations. The observations we consider are simply the word n-grams (up to trigrams) observed in the utterances inside the local context window and the speaker labels associated with these utterances. We annotate each n-gram observation with the associated speaker label, e.g. counselor or client, and the relative position of its host utterance with respect to the current utterance (before or after) so that two identical n-grams associated with different speakers or different contextual positions are treated separately. For instance, if the current utterance is a counselor utterance with the label QUC and includes the bigram “do you”, then we can potentially add the label unigram features `QUC:spk[current]=counselor` and `QUC:ngram[current,counselor]=do_you` to the set of features for the current utterance. Similarly, if the previous utterance is a client utterance with the label FN, then we can potentially add the label bigram features `FN_QUC:spk[current]=counselor` and `FN_QUC:ngram[current,counselor]=do_you`. If the context window includes the previous utterance as well, then we can add features which involve an n-gram and/or speaker label from the previous utterance such as the label bigram features `FN_QUC:spk[before]=client` and `FN_QUC:ngram[before,client]=uh_huh`.

5. Experiments

5.1. Utterance Level MISC Prediction

We used the train subset of the MI Dataset to train linear chain CRF models with different context window sizes for predicting utterance level MISC codes. Table 2 compares the performance of CRF models with different context window sizes in terms of raw label accuracy when the tag set includes all codes in Table 1. The first model has a context window size of 0, which means it only considers the current utterance and the speaker label associated with it. Second model has a context window size of 1, which means it also considers the previous utterance and its speaker label. Similarly, third model has a context window size of 2 and considers the previous two utterances and speakers in addition to the current utterance and speaker. The only observations combined with code bigrams are speaker labels. Word n-grams are only combined with the current code unigram since combining them with code bigrams results in a much larger feature space and significantly slows down the training and decoding for no gain in accuracy. We also noticed that pruning observations that occur only once in the training data significantly improves the training speed without any loss in accuracy. We also experimented with different context sizes for word n-gram and speaker label features (omitted here) but did not see any improvement over the first model in Table 2. It is interesting that contextual features do not improve CRF tagging accuracy in this case.

Table 3 gives a code-by-code performance breakdown of the first model in Table 2 in terms of precision, recall and f1-score. Examining the performance for individual codes, we can see that a large number of MISC codes is hardly ever predicted by the model, which is not surprising given the imbalanced distribution of codes in the training data (Table 1). We address this data sparsity problem by clustering codes. We group all client codes under a single CLI category, retain FA, GI, QUC, QUO, REC and RES as separate categories and group all other counselor codes under the COU category. This clustering is inspired by the code reduction done in MITI coding. The accuracy of our best CRF model (context window size 0) trained on this simpler code set is 87.38%.

Table 2: MISC Code Prediction Accuracy

Context Size	0	1	2
Accuracy	78.99%	78.57%	78.48%

Table 3: Breakdown of MISC Prediction Performance

	Precision	Recall	F1-Score
ADP	0.00	0.00	-
ADW	0.26	0.09	0.14
AF	0.69	0.63	0.66
CO	0.67	0.01	0.03
DI	0.10	0.02	0.03
EC	0.38	0.15	0.22
FA	0.92	0.96	0.94
FI	0.75	0.42	0.53
GI	0.69	0.80	0.74
QUC	0.75	0.69	0.72
QUO	0.82	0.80	0.81
RCP	-	0.00	-
RCW	-	0.00	-
REC	0.47	0.42	0.45
RES	0.47	0.50	0.49
RF	-	0.00	-
ST	0.66	0.33	0.44
SU	0.58	0.17	0.26
WA	0.00	0.00	-
C+	0.00	0.00	-
C-	-	0.00	-
FN	0.86	0.97	0.91
R+	0.39	0.13	0.20
R-	0.47	0.14	0.21
O+	0.18	0.02	0.04
O-	0.28	0.05	0.08
TS+	1.00	0.01	0.03
TS-	-	0.00	-

5.2. Dialog Act Tagging

As an alternative to utterance level MISC prediction, we trained a number of linear chain CRF models on the Switchboard-DAMSL dataset with different context window sizes for predicting dialog acts. Table 4 compares the performance of five of these models on the Switchboard-DAMSL test set in terms of raw label accuracy. The first and second models have context window sizes of 0 and 1 respectively. The third, fourth and fifth models have separate context window sizes for word n-grams (0 for all three models), and speaker labels (1, 2, 3 respectively). Similar to the CRF models for MISC prediction, these models combine only speaker labels with tag bigrams. Word n-grams are combined only with the current tag unigram. Again, observations that occur only once in the training data are pruned. As in the case of MISC coding, we also trained a CRF model (context window size 1) with a simpler tag set consisting of 7 tags as described in [10] that achieves 84.78% accuracy.

Table 4: Dialog Act Tagging Accuracy Results (42 Tags)

Ctx	0-0	1-1	0-1	0-2	0-3
Acc	75.39%	76.25%	76.42%	76.59%	76.59%

5.3. Session Level MITI Prediction

We use the session level tallies of utterance level code and dialog act predictions on our MI test set to predict binarized (High: $\geq$ 4 on Likert scale vs. Low: $<$ 4) session level global ratings assigned to each session as part of MITI coding. Table 5 compares the best CRF models for predicting MISC codes with those for predicting dialog acts. The oracle setup uses reference MISC codes for the test set to predict session level global MITI ratings. The baseline setup assigns all sessions to the majority class (High or Low). MISC28, MISC8, DA42, DA7 refer to our best CRF models trained with full MISC code set, simplified MISC code set, full Switchboard-DAMSL code set and simplified Switchboard-DAMSL code set respectively. It is interesting that the predictions of DA7 outperform the predictions of MISC28 and MISC7 when they are used as features for predicting session level global ratings.

Table 5: Session Level Global Ranking Prediction Results

		Prec.	Recall	F1-Score	Acc.
Empathy	Oracle	0.85	0.85	0.85	85.05%
	MISC28	0.77	0.78	0.75	77.57%
	MISC8	0.75	0.76	0.73	75.70%
	DA42	0.72	0.70	0.71	70.09%
	DA7	0.78	0.79	0.78	78.50%
	Baseline	0.48	0.69	0.57	69.16%
Spirit	Oracle	0.77	0.77	0.77	76.64%
	MISC28	0.69	0.69	0.69	69.16%
	MISC8	0.75	0.74	0.72	73.83%
	DA42	0.72	0.69	0.70	69.16%
	DA7	0.76	0.75	0.75	74.77%
	Baseline	0.37	0.61	0.46	60.75%

6. Conclusion

In this work we presented early results towards extracting information about behavioral codes from the lexical channel using manual session transcripts. We modeled the utterance level MISC coding process as a sequence tagging problem and used linear chain CRF models to predict utterance level codes. We then used session level tallies of code predictions to predict global MITI ratings. As a contrast, we used domain independent dialog act taggers trained on Switchboard-DAMSL to predict utterance level dialog acts and used session level tallies of dialog act predictions to predict global MITI ratings. The results given in Table 5 suggest that domain independent dialog acts can be as effective as highly specialized behavioral codes in terms of predicting therapy quality. It is quite interesting and thought provoking that the reduced DAMSL code set is competitive with the MISC code set in this regard. It can be argued that the lion's share of the lexical signal relevant to overall MI quality (i.e. empathy and spirit) is captured by the general dialogue patterns and 'good therapy' is largely a by-product of a style of communication that is more general to typical conversation, rather than highly specific to psychotherapy.

The fundamental limitation of our approach is its reliance on lexical information derived from session transcripts. Past work [11] has established that the type of lexical information exploited in this work can also be extracted from ASR lattices. Also, it is well known that a subset of dialog acts and in accordance behavioral codes can only be disambiguated with the help of prosodic cues. We plan to address these limitation in our future efforts.

7. References

[1] W. Miller and S. Rollnick, *Motivational Interviewing: Preparing People for Change*. Guilford Press, 2002.

[2] W. R. Miller, T. B. Moyers, D. Ernst, and P. Amrhein, “Manual for the Motivational Interviewing Skill Code (MISC) Version 2.1.” [Online]. Available: http://casaa.unm.edu/download/misc.pdf

[3] B. L. Burke, C. W. Dunn, D. C. Atkins, and J. Phelps, “The emerging evidence base for motivational interviewing: A meta-analytic and qualitative inquiry,” *Journal of Cognitive Psychotherapy*, vol. 18, pp. 309–322, 2005.

[4] E. Proctor, H. Silmere, R. Raghavan, P. Hovmand, G. Aarons, A. Bunger, R. Griffey, and M. Hensley, “Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda.” *Adm Policy Ment Health*, vol. 38, no. 2, pp. 65–76, 2011.

[5] T. B. Moyers, T. Martin, J. Manuel, and W. R. Miller, “The Motivational Interviewing Treatment Integrity (MITI) Code: Version 2.0.” [Online]. Available: http://casaa.unm.edu/download/miti.pdf

[6] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van, and E. dykema Marie Meteer, “Dialogue act modeling for automatic tagging and recognition of conversational speech,” *Computational Linguistics*, vol. 26, pp. 339–373, 2000.

[7] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML ’01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289. [Online]. Available: http://dl.acm.org/citation.cfm?id=645530.655813

[8] 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.

[9] 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.

[10] E. Shriberg, A. Stolcke, D. Jurafsky, N. Coccaro, M. Meteer, R. Bates, P. Taylor, K. Ries, R. Martin, and C. Van Ess-Dykema, “Can prosody aid the automatic classification of dialog acts in conversational speech?” *Language and speech*, vol. 41, no. 3-4, pp. 443–492, 1998.

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