RUNNING HEAD: REFLECTION DETECTION

"It sounds like…": A Natural Language Processing Approach to Detecting Counselor Reflections in Motivational Interviewing

Doğan Can [a], Rebeca A. Marín [b], Panayiotis G. Georgiou [a], Zac E. Imel[c],

David C. Atkins [b], & Shrikanth S. Narayanan [a]

[a] University of Southern California
[b] University of Washington
[c] University of Utah

Author Note. Funding for the preparation of this manuscript was provided by the National Institutes of Health / National Institute on Alcohol Abuse and Alcoholism (NIAAA) under award number R01/AA018673. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The methodology and results were previously presented at the following venues: 1) INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, 2) ABCT 2013, 47th Annual Convention of the Association for Behavioral and Cognitive Therapies.

Please direct all correspondence concerning this manuscript to:

Doğan Can
University of Southern California
3710 S. McClintock Ave, RTH 320
Los Angeles, CA 90089
PHONE: (213) 740-3477 | FAX: (213) 740-4651
Email: dogancan@usc.edu

Abstract

The dissemination and evaluation of evidence based behavioral treatments for substance abuse problems rely on the evaluation of counselor interventions. In Motivational Interviewing (MI), a treatment that directs the therapist to utilize a particular linguistic style, proficiency is assessed via behavioral coding - a time consuming, non-technological approach. Natural language processing techniques have the potential to scale up the evaluation of behavioral treatments like MI. We present a novel computational approach to assessing components of MI, focusing on one specific counselor behavior – reflections – that are believed to be a critical MI ingredient. Using 57 sessions from 3 MI clinical trials, we automatically detected counselor reflections in a Maximum Entropy Markov Modeling framework using the raw linguistic data derived from session transcripts. We achieved 93% recall, 90% specificity, and 73% precision. Results provide insight into the linguistic information used by coders to make ratings and demonstrate the feasibility of new computational approaches to scaling up the evaluation of behavioral treatments.

Keywords: Motivational Interviewing, Fidelity Assessment, Dissemination, Natural Language Processing

"It sounds like…": A Natural Language Processing Approach to Detecting Counselor Reflections in Motivational Interviewing

Motivational Interviewing (MI; Miller & Rollnick, 2013) is a widely studied cognitive behavioral treatment with parallel roots in humanistic counseling traditions such that counselor fidelity to MI is characterized by direct and strategic elicitation of client talk about changing the target behavior coupled with a non-judgmental, empathic interpersonal style. Although the efficacy of MI has been established for a number of problems (Miller & Rose, 2009), research on the therapist behaviors that lead to positive treatment outcomes are inconclusive (Huebner & Tonigan, 2007; Magill et al., 2014). Developing a clearer understanding of the active ingredients of MI (and other behavioral treatments) could lead to better training protocols, more effective dissemination, and ultimately revisions and extensions of the treatment itself (Nock, 2007).

Unfortunately, conducting research on in-session behaviors in MI is extremely difficult due to the nature of audio or video recordings of counselor-client discourse (Longabaugh, 2007). This rich but highly unstructured data is typically quantified by applying a behavioral coding system (McKay, 2007). In MI, the Motivational Interviewing Skills Code (MISC; Miller, Moyers, Ernst, & Amrhein, 2008) is the gold-standard behavioral coding system. The MISC and other treatment fidelity measures (see Moyers, Martin, Manuel, Miller, & Ernst, 2010) have revealed important theory-driven relationships between counselor and client behavior in MI therapy (Magill et al., 2014). However, similar to other behavioral coding methods, the MISC has a number of disadvantages. The first is the expense of coding in both time and money. Moyers et al. (2005) noted an average of 85 minutes of coding time per each 20 minute segment of a MI session, or roughly 4:1, and in a separate study, a slightly higher figure of 90-120

minutes per 20 minute segment, or 5-6:1 (Moyers et al., 2005). Although software enhancements can improve coding time to some extent (see, e.g., Glynn, Hallgren, Houck, & Moyers, 2012), detailed behavioral coding is time-consuming. The costs of behavioral coding limit its application in number and length of sessions coded; for example, the Magill et al. (2014) meta-analysis included a total of 783 sessions across all studies. Secondly, establishing reliability can be problematic given the complexity of coder training which requires roughly 40 hours of initial training time in addition to any further training due to coder drift and / or turnover (Miller et al., 2007; Moyers et al., 2005). Finally, beyond the one-time start-up costs of establishing the coding team, coding scales linearly (i.e., coding 100 tapes is approximately 10 times the work of coding 10 tapes). Accordingly, ongoing human coding does not "scale up" to larger tasks, limiting both the pace and scope of scientific research on behavioral treatments. Research on the active ingredients involved in MI could benefit from an alternative methodology that sidesteps the costs and reliability problems associated with manual behavioral coding. This paper is a case study on automating behavioral coding of MI sessions.

A computer-based tool that could use raw text from an MI session transcript to predict human coding could avoid a major bottleneck in MI mechanism research, that is the dependency on the amount of human effort required for behavioral coding. Such a system can code 10 vs. 1,000 sessions with no extra cost other than the additional computational resources, which are insignificant compared to the cost of human coding, and could support and extend human evaluation of MI sessions by significantly reducing the time spent for behavioral coding. With such a system, typical human coding concerns like establishing coder reliability, preventing coder drift and training new human coders would practically disappear given the system

performs reliably compared to a gold-standard coding reference set. Moreover, the models and algorithms can be improved and extended over time, and the coding procedure can easily be repeated in the event of such refinements. Maybe even more importantly such a system can support reproducible results.

Behavioral Signal Processing (BSP; Narayanan & Georgiou, 2013) is an emerging inter-disciplinary research area that is perfectly suited to studying and developing such computer-based tools. BSP aims to model and give insights into human behavior by quantifying complex behavioral constructs based on low-level, observed behavioral cues. It combines computational techniques from multimodal signal processing (e.g., audio, visual, physiological signals), natural language processing, and machine learning with psychology and behavioral science expertise to produce computational tools that support and extend the capacity of human evaluation. In the case of MI, the MISC system represents a phenomenally labor-intensive task, and thus, computational approaches within BSP – such as speech and natural language processing (NLP) techniques – could offer avenues for streamlining and automating the coding process (Crammer et al. 2007, Georgiou et al., 2011, Goldstein et al., 2007).

As a first step towards automated behavioral coding of MI sessions, in this paper we turn our attention to detecting a core class of counselor responses in MI, *reflections*. Specifically, reflections are a core MI micro-skill used as a form of hypothesis testing to determine the meaning of client statements. MI describes two types of reflections – simple and complex. Simple reflections repeat or rephrase the client's statements and are intended to help the counselor comprehend the client's understanding of their situation. Complex reflections have an added piece of interpreting the client's statement either by using new words or guessing at the

client's meaning (Miller & Rollnick, 2013). Reflections are thought to be a powerful way of expressing empathy and building rapport. They can be used to understand the client's viewpoint, address resistance, and highlight discrepancies between the client's values and their target behavior (Ginter & Choate, 2003). Moreover, counselor use of reflections has been consistently linked to behavior change across a variety of populations and behaviors, including cannabis use (McCambridge, Day, Thomas, & Strang, 2011), adherence to HIV anti-retroviral therapy (Thrasher, Golin, Earp, Tien, Porter, & Howie, 2006), and alcohol (e.g., Tollison et al., 2008).

Natural Language Processing (NLP) as a Method for Quantifying Textual Information Derived from Counselor-Client Interactions in Motivational Interviewing

Designing a statistical model to detect reflections from text (e.g., session transcripts) presents a significant challenge due to the potential number of predictor variables. Depending on the speech rate, a 50-minute MI session transcript will have approximately 12,000 – 15,000 words. Thus, even a simple model that includes nothing but a single indicator variable for each word observed in the session transcripts will end up with thousands of potential predictors.

We can reduce the overall complexity of the data by extracting linguistic "features" (i.e., predictors) from transcripts that map directly onto patterns of speaking that counselors may use when making reflections. Feature extraction is a general term in machine learning literature that refers to a form of dimensionality reduction where input data (e.g., transcripts) are transformed into a reduced representation composed of features that characterize the construct of interest. For instance, we can represent each utterance in a session transcript as a set of words contained in that utterance and discard the sequence information contained in regular text. This approach to modeling textual data is called the "bag of words" approach in NLP parlance. At a high level,

feature extraction is analogous to what humans do in behavioral coding; for example, to assign a behavioral code to a spoken utterance, they make use of the information conveyed by the specific language use. In the present work, we draw on NLP methods and focus on three specific linguistic features that may characterize reflections. We settled on these features by carefully considering what kinds of utterances are typically coded as reflections and how reflections are described in the MISC manual (Miller et al., 2008). We provide a brief overview here, and then describe the specific application in the present study in the Methods section.

N-grams. Counselors may use specific word sequences when making reflections. These are called n-grams in NLP parlance. They consist of contiguous sequences of n words (e.g., a uni-gram is a single word, a bi-gram is a two word sequence, tri-gram is a three word sequence). For instance, a counselor might use the phrases "from what I gather" and "it sounds like" in reflective statements. These phrases may be even more common in complex reflections than in simple ones, since most analogies, metaphors, similes, exaggerations, and summaries fall under the complex reflections category (Miller et al., 2008). In addition to the n-grams used in a target utterance, i.e. the counselor utterance we would like to classify as reflection or not, n-grams used in the local conversational context of that utterance may also be informative in detecting reflections. We call these contextual n-grams. For instance, client utterances following a reflection may include confirmations or refutations such as "yeah" and "not really".

Meta-Features. Reflections occur during a conversational interaction, which means they not only affect the dialog flow, e.g. clients tend to confirm a reflection with statements like "yeah", but also are affected by what is happening in the local context. From this perspective, a reflection could potentially be inferred by examining other information collected from the local

conversational context surrounding a target counselor utterance such as whether the previous utterance was a client utterance or whether we predicted another reflection occurrence in a prior or subsequent counselor talk turn. For example, given that an individual talk turn may contain one or more utterances (and MI behavioral codes are codes for utterances), reflections might be more likely to occur immediately after a client utterance as opposed to a counselor utterance, or they might occur in succession in a counselor talk turn.

Client-Counselor Similarity. A reflection's primary function is to "capture and return" the clients meaning. Especially in the case of simple reflections, counselors may simply repeat or mildly rephrase what the client has said. Consequently, the similarity of n-grams between a counselor utterance and the preceding client utterances may be relevant to detecting reflections with particularly high similarity indicating a reflection.

Current Study

The present study is an initial, "proof of concept" study exploring whether NLP techniques such as those described above could be used to build a statistical model for a specific counselor intervention that is central to an established evidenced based treatment, reflections. The primary focus is on the accuracy with which an NLP model can automatically detect human generated reflection codes. Furthermore, we are also interested in which of the linguistic components (i.e., n-grams vs. contextual n-grams vs. meta-features) contributed the most to predicting reflections, which could be informative about the process of making reflections.

Method

Data Source

The data for the present study came from three different MI intervention studies. The first MI study recruited from community primary care clinics where many of the clients were polysubstance users (Roy-Byrne et al., 2014). The other two studies involved event-specific interventions for college students' drinking focused either on 21st birthdays (Neighbors et al., 2012) or spring break trips (Lee et al., 2014). The present MISC-coded dataset included 57 unique sessions. Some of the sessions were coded by multiple coders in order to establish coder reliability. As a result there were a total of 108 coded sessions. Session lengths ranged from 15 to 45 minutes. All studies were based in Seattle, Washington, and all original trial methods were approved by the University of Washington Institutional Review Board.

Transcription and Pre-processing

Each session was professionally transcribed following strict guidelines for denoting specific speech patterns (e.g., speech-overlaps, disfluencies, interruptions, repetitions, etc.). Prior to MISC coding, we normalized each transcript by removing word-external punctuation symbols (i.e., hyphens, apostrophes and underscores were retained), splitting utterances into words in the Penn Tree Bank style (i.e., utterances were split on word boundaries and apostrophes; Marcus et al., 1993), and finally converting all words to lowercase.

Measures

Motivational Interviewing Skills Code (MISC; Miller et al., 2008). The MISC Version 2.1 was applied as a mutually exclusive and collectively exhaustive coding system, where every utterance was assigned a single behavior code. Two trainers who are part of the Motivational Interviewing Network of Trainers with prior experience in training coding teams trained three undergraduate and post-baccalaureate students. Eleven session recordings were used to train

coders, including real and standardized patient recordings. After training was completed, the team established initial inter-rater reliability by triple-coding twelve MI sessions. Based on intra-class correlation coefficients (ICCs; Shrout & Fliess, 1979), coders obtained .50 or better agreements on most individual codes (72%), and 34% of codes were at or above .75 agreement. ICCs were .95 and .86 for simple and complex reflections, respectively.

Not all code categories are equally represented in our dataset. To deal with sparsity problems caused by this uneven distribution and the limited size of our dataset, all client codes were combined into a single, meta-code "CL," and all counselor codes – except reflections – were combined into a meta-code "CO." Finally, reflections (both simple and complex) were treated as a single code "RE." After these reductions, we had a total of 34,388 CL codes, 26,807 CO codes, and 7,115 RE codes (4084 simple + 3031 complex) in our dataset.

Linguistic Feature Extraction

Figure 1 shows a small excerpt of an MI transcript (comprised of four utterances in three talk turns), which will be used to explain feature extraction and encoding.

N-grams. Counselors tend to use certain linguistic constructs while reflecting. Phrases like "from what I gather" and "it sounds like" are common in utterances that are reflections. Furthermore, other types of counselor and client categories tend to have their own specialized language patterns (e.g. question, giving information, reason). We use word n-gram (word sequence) features automatically extracted from the current utterance, i.e. no predefined list of n-grams is provided, to capture particular language use. For instance, the n-gram features extracted from the third utterance in Figure 1 (and indicated by the red arrow) include the uni-grams "you" and "like", the bi-gram "you are", and the four-gram "it would be like," among others.

Contextual N-grams. A typical MI scenario is when the client responds to a reflection with a confirmation. Many client utterances following a reflection statement start with the discourse marker “yeah”. This behavior is not distinguished by MISC and is usually coded as a *Follow/Neutral* (FN) statement (a general code that indicates the client response is not indicative of client change talk). We use word n-gram features extracted from utterances in the local context of current counselor utterance to capture this kind of phenomenon. We prefix each n-gram feature with the associated speaker (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 the contextual n-grams for the second utterance in Figure 1 (highlighted in blue) include those extracted from the previous client utterance, e.g. “I would not”, from the previous counselor utterance, e.g. “you do not want”, and from the next client utterance, e.g. “yeah”.

Meta-Features. Contextual speaker and code information can be helpful in discriminating reflections from other counselor speech. We extracted speaker and code features from the local context and used them as another source of information. During model training, we used the actual (human generated) codes associated with each contextual utterance to learn the language and code occurrence patterns, while during evaluation we replace those with the codes automatically generated by our system. For instance the meta-features extracted for the third utterance in Figure 1 (highlighted in red) include “Counselor:Client” (current utterance is in between a counselor utterance and a client utterance), and “CL_RE:Client” (previous two utterances were coded as CL and RE respectively; the next utterance is a client utterance).

Similarity Features. The primary function of reflections is to "capture and return" the client talk. Especially in the case of simple reflections, counselors tend to simply repeat or mildly rephrase what the client has said before. With that in mind, we gathered all word stem n-grams shared between a counselor utterance and the preceding client utterances in the local context, and then eliminated those which do not include at least one content word, i.e. a noun, verb, adjective or adverb. We used the remaining word-stem n-grams and the n-grams of corresponding part-of-speech tags as binary similarity features. For instance the similarity n-grams extracted for the second utterance in Figure 1 (shown in purple) include word n-gram "here" and the corresponding part-of-speech n-gram "ADVERB".

Prediction Model and Evaluation

We treated the utterance-level MISC coding process as a specialized form of dialog act tagging, a standard NLP task (Stolcke et al., 2000; Sridhar et al., 2009). Dialog act taggers are sequence predictors that process input utterances in time order and predict a dialog act tag for each utterance based on the features extracted from a local context around the utterance. A "tag" is simply a type of code that is particular to linguistic exchanges (e.g., "question" or "statement" could be tags in a dialog act tagging study). In this study, we limited our attention to the local utterance context and evaluated a tagging system that uses automatically extracted linguistic features with rich contextual information to detect reflection occurrences.

Feature Selection. Using all word n-gram features extracted from a large local context size (e.g., using the n-grams in utterances 7, 8, or 9 turns away from the current utterance) would create a massive set of predictors, which is not feasible given the size of the data. We experimented with various combinations of features parameterized by factors such as the context

size used for extracting contextual features (e.g. how many talk turns were used to define the local context), which can be different for each type of contextual feature, and the maximum n-gram sizes, which again can be different for the current and the contextual utterances. Also, part of the modeling process was using feature selection to limit the word n-gram features to the most informative ones as determined by Shannon's (1948) entropy measure, a measure of our uncertainty about our source of information, e.g. a particular n-gram feature.

Maximum Entropy Markov Model (MEMM; McCallum et al., 2000). Thus far, we have described NLP methods for computing various quantities from text (i.e., the classes of linguistic features), but to use these as predictors of reflections, it is necessary to connect the features with the outcome code classes in an appropriate statistical model. Since reflections occur in a dynamic context of interaction, the model we desire should be able to capture this temporal context. MEMMs are discriminative sequence models that directly model the conditional distribution of each hidden (or latent) state (i.e., code class such as CL, CO, or RE), given a set of predictors using logistic regression. They are used extensively in sequence prediction tasks (McCallum et al., 2000). MEMMs are very flexible when it comes to the types of features that can be included in the model, which is a limitation of other popular generative sequence models like the Hidden Markov Model (McCallum et al., 2000).[1] For instance, arbitrary features of observations, which are not required to be statistically independent of each other, can be used as predictors, allowing domain-specific knowledge to be easily inserted into the model.

[1] We also experimented with a Hidden Markov Model (HMM), but the performance was inferior to the MEMM described above. HMMs have certain shortcomings when they are used for prediction since they are not designed for discriminating between observations and have limitations on the types of features that can be included when used for prediction. Specifically, in the present application contextual n-grams cannot be included in HMMs.

Figure 2 presents a graphical representation of the process that the MEMM uses in making decisions about individual utterances. The MEMM uses the features described earlier (and shown in Figure 1) as input predictors to a sequence of logistic regression models predicting the outcome codes[2] (i.e., CO or RE) and searches for the code sequence with the best performance using the Viterbi algorithm (McCallum et al., 2000; see Figure 2 caption). There are two key differences of a MEMM from a standard logistic regression model (i.e., over and above the fact that it is using linguistic features as predictors). First, the model works in a sequential fashion, starting with the first utterance and moving through the code sequence to the final utterance. Second, the MEMM does not make a final decision about each individual utterance, but instead makes decisions across a sliding window of utterances (in our case the 2 utterances before and subsequent to the reference utterance). Thus, it will make a preliminary (probability) estimate of whether a given utterance is a reflection, but this estimate may be updated (i.e., changed) when the model considers a later utterance. In Figure 2, the MEMM initially proposed a CO code for utterance 2, which was subsequently updated to be RE, and the final assignment of codes is seen in the yellow path of arrows. Some meta-features used by our model (e.g. the code assigned to previous counselor utterance) depend on the code predictions made for earlier utterances in a session. One way to include such features in prediction is to follow a 'greedy decision process', where utterances are processed in order and earlier code predictions are fixed while processing a new utterance. However, greedy decisions result in the propagation of prediction errors from one utterance to the next. The Viterbi algorithm does not commit to

[2] Note that client utterances are automatically assigned the code CL in our model.

decisions for earlier codes but instead stores likely code predictions for earlier utterances in a dynamic decoding graph and searches for the best sequence of codes through this graph.

Prediction Accuracy and Cross-Validation. The key result is the accuracy with which the MEMM can predict the presence of a therapist reflection (vs. any other therapist). We report precision (number of true positives divided by the number of hypothesized positives, e.g. if the system predicts 100 reflections and 70 of them are actually reflections then the precision would be 70%), recall (sensitivity; number of true positives divided by the number of reference positives, e.g. if there are 100 reflections in a session and the system predicts 80 of them as reflections then the recall would be 80%), specificity (number of true negatives divided by the number reference negatives, e.g. if there 100 utterances in a session that are not reflections and the system predicts 90 of them as non-reflections then the specificity would be 90%) and F-score (harmonic mean of precision and recall) for each feature combination investigated. We use F-score as our primary metric for assessing a model's accuracy. F-score is a balanced measure of system performance since it accounts for both how good the system is at detecting reflections (high recall) as well as not predicting non-reflections as reflections (high precision). It takes a value between 0% and 100% with higher numbers indicating better performance.

As is typical in machine learning applications, the accuracy numbers given in this paper are based on cross-validation. Given the relatively small sample size, we used "leave one out" cross-validation (LOOCV; Kuhn & Johnson, 2013). In LOOCV, the statistical model is fit to all sessions except one, and its accuracy is evaluated on the one session not included in the model fit. This is then iterated for every session in the data and the presented accuracy results are an average over all the individual sessions. If the excluded session was coded multiple times, then

we leave all copies of that session out of the model to not cause any dependencies in data, i.e. all of our results are average numbers obtained from 57-fold cross validation.

Results

Table 1 summarizes the reflection detection performance for each feature combination investigated. Our first model uses only the similarity features, i.e. shared n-grams, extracted from the local context (up to 4-grams, context window includes last 9 client utterances). This model is able to detect reflection occurrences at an acceptable precision level of 67%; yet the recall is quite poor at 31% (first row in Table 1). Our second model uses only the n-grams (up to 4-grams) from the current utterance. In spite of its simplicity, it provides a fairly competitive baseline for detecting the language use specific to reflections with an F-score of 73% (second row in Table 2). When we include the contextual n-grams (up to 4-grams) from the immediate utterance context (previous and next utterances), the F-score value increases to over 80% (compare second and third rows in Table 2), which indicates that there is rich information in the local neighborhood of an utterance that helps discriminate between reflections and other counselor interventions. Adding meta-features further improves the precision to 72% and F-score to above 81% (fourth row in Table 2). This gain is largely due to the predicted code n-grams that contextualize the current utterance within the dialogue. Finally, we also tried combining the similarity and the n-gram features (fifth row in Table 2), which resulted in a slight decrease in performance. We should note that the F-score values for all models but the first one are significantly higher than chance values. As a comparison, a system predicting all therapist utterances as reflections would achieve 21% precision, 100% recall and 35% F-score, while a

system randomly guessing therapist utterances as reflections as often as they are observed in the data (21%) would on average achieve 21% precision, 21% recall and 21% F-score.

Discussion

We presented a method for automatically detecting counselor reflections in Motivational Interviewing that compared several sources of information drawn from session transcripts. Model performance was quite strong, suggesting that NLP approaches hold promise in scaling up the evaluation of provider fidelity in evidence-based psychotherapies like MI.

The current results complement other, recent approaches to using modern NLP and text-based machine learning approaches to modeling provider fidelity ratings (Atkins et al., 2014; Gallo et al., 2014). Relative to these approaches, the current method is unique in that it bases its prediction on the local context of an utterance, in addition to the utterance being predicted. One of the key tasks moving forward will be comparing and contrasting different NLP and text-based machine learning methods to learn which methods are optimal for which codes. Reflections, by their nature, are related to the local context and have particular, salient words and phrases. However, this is not true of all provider fidelity codes. As one example, MI coding systems rate empathy for an entire session, not on a per-utterance basis, precisely because MI views empathy as an overall, gestalt experience within a session. Thus, a priori, we might expect that the current NLP methodology for reflections would be a poor fit for predicting empathy. Future research will need to make precisely these types of comparisons, matching methods to text and codes.

One advantage of the model tested here is that results may provide further insight into how specific treatment components are evaluated. For example, there appears to be substantial information about the coding of a standard therapist intervention like reflections in the speaking style of the counselor (e.g., the n-gram features extracted from the target counselor utterance).

While this finding might be attributed to the common language patterns counselors use during reflections ("it sounds like"), it can also point to the human process of perceiving a reflection, i.e. coders might be influenced as much by how something is said as they are by what is being said. The n-gram effect may also have implications for extending NLP methods to other behavioral treatments that may be more "content" driven (as opposed to MI which directs a therapist to have a particular linguistic style). For example, in an intervention like Prolonged Exposure, there may be particular moments in the session when a therapist asks for ratings of subjective units of distress that might be captured by n-gram based models similar to the one implemented here.

However, the importance of language style in the prediction of reflections also raises questions about whether the accuracy of a counselor reflection drives a coder's decision to label an utterance a reflection. Essentially, a coder might be tagging the form of a reflection, but not its actual content. It is interesting to note the rather poor performance of similarity features. This result might be due to our simplistic approach to modeling utterance similarity. Our model relies on detecting shared n-grams between a counselor utterance and the prior client speech for measuring similarity. A more involved approach based on semantic similarity rather than lexical similarity might prove to be more successful. Furthermore, it may be that other counselor talk (besides reflections) also shares content words with prior client speech (e.g., both a client and therapist are talking about problems with drinking), creating a ceiling effect wherein similarity is generally high across many utterance pairs.

Another important observation is the effect of including contextual n-grams into the model. Contextual n-gram features extracted from immediate context of the target utterance significantly increases the number of reflections detected by our system with a relatively small

loss in detection precision. An important component of context was the client response to a reflection – e.g., when the client said "yeah" after a counselor utterance. Here it may be that coders were looking for confirmation of an accurate reflection in the response of the client.

Limitations

There are several limitations of our approach, the most notable being the reliance on human generation of the transcripts. While the need for humans to transcribe treatment sessions limits the potential for NLP methods to evaluate large numbers of sessions, advances in automatic speech recognition will decrease the need for human transcription in the future. In addition, there are likely other indicators of human coded reflections that are not found in transcripts. For example, there may be contextual indicators such as the tone of the previous client utterance, which might provide an opportunity for a reflection. Our approach clusters both types of reflections into a single code and does not discriminate between simple and complex reflections. While the main motivation behind this clustering was to make sure we had enough data for fitting the model, it also makes the prediction problem easier since these two codes are similar in that therapists tend to use similar constructs in both kinds of reflective statements. Given that human coders routinely confuse these two codes and many therapist utterances fall somewhere between simple and complex reflections, discriminating between these two sub-codes is likely a harder task than discriminating reflections from other therapist speech. Finally, this paper limits its attention to reflections but the methods described can be easily adapted to detect other codes from session transcripts.

Conclusions

A reliance on human based behavioral coding is a major impediment to scaling up the evaluation of evidence based behavioral treatments. The n-gram MEMM model tested in this paper provides strong initial evidence that NLP methods may provide one method for exploring the content of psychotherapy in a way that does not rely on humans.

References

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*, 49. doi: 10.1186/1748-5908-9-49

Crammer, K., Dredze, M., Ganchev, K., Talukdar, P. P., & Carroll, S. (2007, June). Automatic code assignment to medical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing* (pp. 129-136). Association for Computational Linguistics.

Gallo, C., Pantin, H., Villamar, J., Prado, G., Tapia, M., Ogihara, M., Cruden, G., & Brown, C. H. (2014). Blending qualitative and computational linguistics methods for fidelity assessment: Experience with the Familias Unidas preventative Intervention. *Administration and Policy in Mental Health,* doi: 10.1007/s10499-014-0538-4

Georgiou, P. G., Black, M. P., Lammert, A. C., Baucom, B. R., & Narayanan, S. S. (2011). "That's Aggravating, Very Aggravating": Is It Possible to Classify Behaviors in Couple Interactions Using Automatically Derived Lexical Features?. *Affective Computing and Intelligent Interaction* (pp. 87-96). Springer Berlin Heidelberg.

Glynn, L. H., Hallgren, K. A., Houck, J. M., & Moyers, T. B. (2012). CACTI: Free, open-source software for the sequential coding of behavioral interactions. *PLoS ONE, 7*(7), e39740.

Goldstein, I., Arzumtsyan, A., & Uzuner, Ö. (2007). Three approaches to automatic assignment of ICD-9-CM codes to radiology reports. *AMIA Annual Symposium Proceedings* (Vol. 2007, p. 279). American Medical Informatics Association.

Huebner, R. B., & Tonigan, J. S. (2007). The search for mechanisms of behavior change in evidence-based behavioral treatments for alcohol use disorders: Overview. *Alcoholism: Clinical and Experimental Research, 31*, 1S-3S. doi: 10.1111/j.1530-0277.2007.00487.x

Krupski, A., Josech, J., Dunn, C., Donovan, D., Bumgardner, K., Lord, S. P., . . . Roy-Byrne, P. (2012). Testing the effects of brief intervention in primary care for problem drug use in a randomized controlled trial: Rationale, design, and methods. *Addiction Science & Clinical Practice, 7*(1), 27. doi:10.1186/1940-0640-7-27

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. New York: Springer.

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*(2), 189.

Longabaugh, R. (2007). The search for mechanisms of change in behavioral treatments for alcohol use disorders: A commentary. *Alcoholism: Clinical and Experimental Research, 31*, 21S-32S. doi: 0.1111/j.1530-0277.2007.00490.x

McKay, J. R. (2007). Lessons learned from psychotherapy research. . *Alcoholism: Clinical and Experimental Research, 31*, 48S-54S. doi: 10.1111/j.1530-0277.2007.00493.x

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, Online First Publication, May 19, 2014*. doi: 10.1037/a0036833

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*,*19*(2), 313-330.

McCallum, A., Freitag, D., & Pereira, F. C. (2000, June). Maximum Entropy Markov Models for Information Extraction and Segmentation. *International Conference on Machine Learning* (*ICML)* (pp. 591-598).

McCambridge, J. Day, M., Thomas, B. A., & Strang, J. (2011). Fidelity to Motivational interviewing and subsequent cannabis cessation among adolescents. *Addictive Behaviors, 36*, 749-754. Doi: 10.1016/j.addbeh.2011.03.002

Miller, W. R., Moyers, T. B., Ernst, D. B., & Amrhein, P. C. (2008). *Manual for the Motivational Interviewing Skill Code (MISC), Version 2.1*. New Mexico: Center on Alcoholism, Substance Abuse, and Addictions, The University of New Mexico.

Miller, W. R., & Rollnick, S. (2013). *Motivational Interviewing: Helping People Change*. (3rd ed.). New York, NY: Guilford Press.

Moyers, T., Martin, T., Catley, D., Harris, K. J., & Ahluwalia, J. S. (2005). Assessing the integrity of motivational interviewing interventions: Reliability of the motivational interviewing skills code. *Behavioural and Cognitive Psychotherapy*, *31*, 77–184.

Moyers, T. B., Martin, T., Manuel, J. K., Miller, W. R., & Ernst, D. (2010). Revised global scales: Motivational interviewing treatment integrity 3.1. 1 (MITI 3.1. 1).

Narayanan, S. & Georgiou, P. G. (2013). Behavioral Signal Processing: Deriving Human Behavioral Informatics from Speech and Language. *Proc. of IEEE, 101*(5), 1203-1233.

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*(5), 850-862. doi: 10.1037/a0014386

Nock, M. K. (2007). Conceptual and design essentials for evaluating mechanism of change. *Alcoholism: Clinical and Experimental Research, 31*, 3S-12S.

Roy-Byrne, P., Bumgardner, K., Krupski, A., Dunn, C., Ries, R., Donovan, D., West, I. I., Maynard, C., Atkins, D. C., Cook, M., Joesch, J., M., & Zarkin, G. A. (2014). Brief Intervention for Problem Drug Use in Safety-Net Primary Care Settings. *JAMA*, *312*, 482-501.

Shannon, Claude E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal,* 27(3), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Sridhar, V. K. R., Bangalore, S., & Narayanan, S. (2009). Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. *Computer Speech & Language*, *23*(4), 407-422.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., ... & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, *26*(3), 339-373.

Thrasher, A. D., Golin, C. E., Earp, J. A. L., Tien, H., Porter, C., & Howie, L. (2006). Motivational interviewing to support antiretroviral therapy adherence: The role of quality counseling. *Patient Education and Counseling, 62*, 64-71.

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*(2), 183-194.

Table 1. *Results*

Included Feature Types	Recall (%)	Specificity (%)	Precision (%)	F-Score (%)
Similarity	31.1	97.6	67.1	42.5
N-gram	63.6	97.9	86.7	73.4
N-gram + Context N-gram	94.1	89.3	69.9	80.2
N-gram + Context N-gram + Meta	92.5	89.8	72.7	81.4
N-gram + Context N-gram + Sim.	91.5	90.3	69.6	79.0

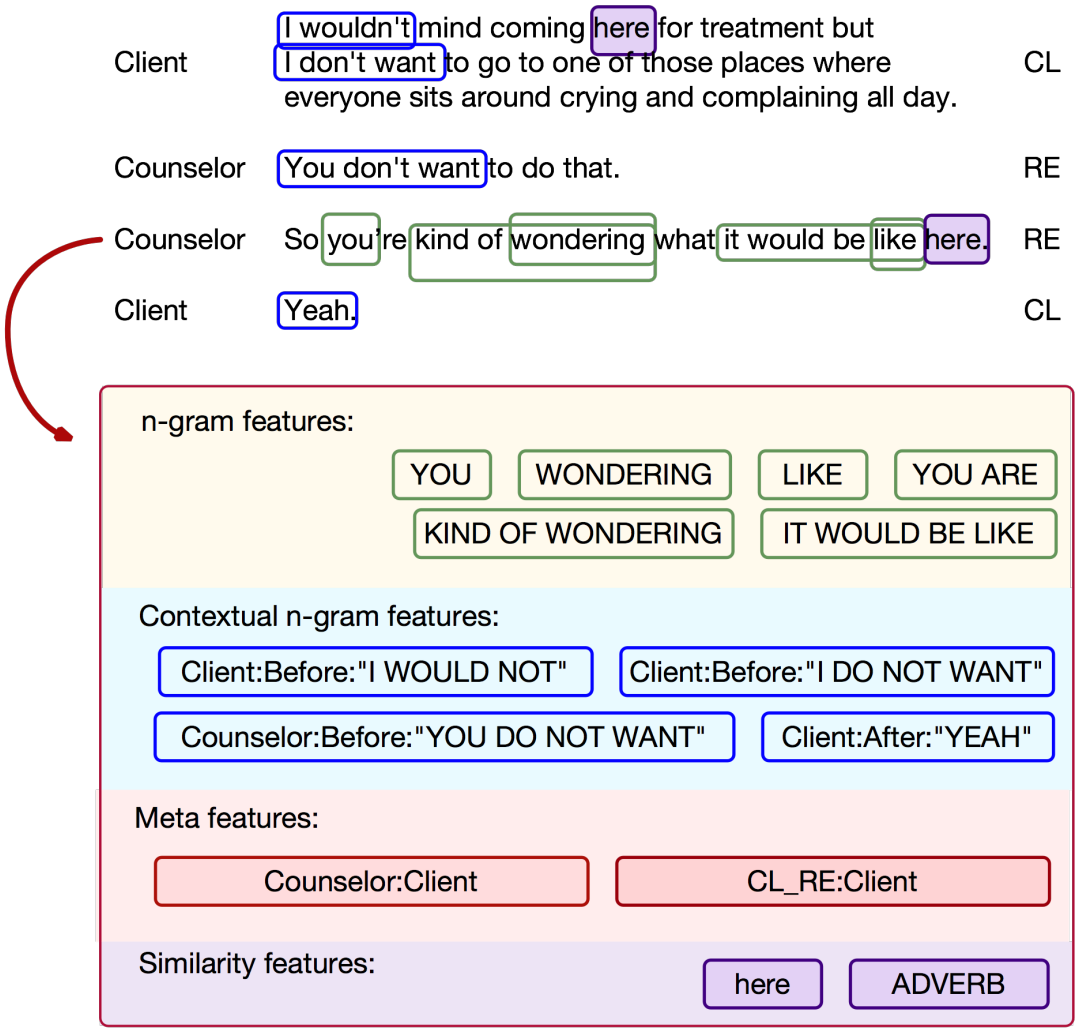

Figure 1. An example dialog snippet that highlights the coded utterance as well as the local context. Each type of n-gram feature evaluated by the MEMM model is highlighted in different colors. N-grams of the coded utterance are highlighted in green, contextual n-grams in blue, meta-features in red, and similarity features in purple.

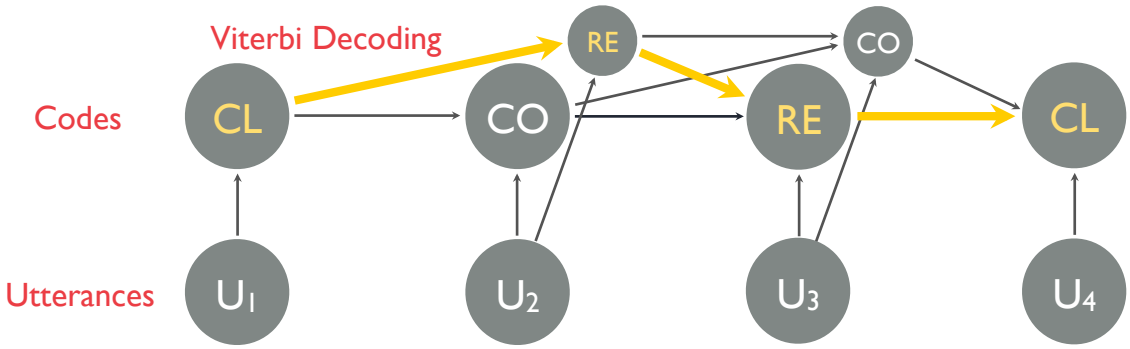

Figure 2. A graphical representation of the Viterbi decoding process for the sample dialog snippet given in Figure 1. In this example, the Maximum Entropy Markov Model (MEMM) can be thought as a sequence of four logistic regression models: 1) client model, 2) counselor model, 3) counselor model, and 4) client model. The first model uses the client label (given in the transcript) for the first utterance, U_1, to assign the code CL, as there is no alternative client code in our setting. Then the second model uses the decision made for U_1 (i.e., CL) as a binary feature in predicting the code for U_2 (i.e., CO or RE) along with other features extracted for U_2 (n-grams, etc.), assigning probabilities to each decision for the second utterance (e.g. CO/0.6, RE/0.4).

Similarly, the third model uses the decisions made for the first two utterances, U_1 and U_2, as binary features in predicting the code for U_3. Finally the fourth model assigns the code CL to the last utterance, U_4, and the Viterbi algorithm chooses the most likely sequence of codes for the entire dialog (provided in yellow). The Viterbi algorithm allows the model to efficiently evaluate alternative code sequences for earlier utterances that are less likely in isolation but may turn out to be more likely when considered in the context of later utterances. Note that the code for U_2 was first CO (i.e., not a reflection), but was later updated to be RE based on the information from U_3 and U_4.