Towards a Computational Framework for Automating Substance Use Counseling with Virtual Agents

Stefan Olafsson
Northeastern University
Boston, MA
olafsson.s@northeastern.edu

Byron Wallace
Northeastern University
Boston, MA
b.wallace@northeastern.edu

Timothy Bickmore
Northeastern University
Boston, MA
t.bickmore@northeastern.edu

ABSTRACT

Motivational interviewing is a counseling technique that involves the in-depth exploration of a person's reasons for and against changing their behavior, and is particularly effective for substance use counseling. We are developing a computational framework that uses techniques from motivational interviewing to conduct substance use counseling sessions by simulating face-to-face interactions with a virtual agent. We evaluated the feasibility of using a virtual agent system that uses a constrained-input modality and dialogue trees to automate parts of motivational interviewing, and report the results conducted with patients at two substance use treatment facilities. We are extending this prototype to encompass all of motivational interviewing by processing information from unconstrained user speech. To that end, we report results from training a dialog act prediction model on 132 transcripts of patient-provider counseling sessions. Our best model realized an F1 score of 0.62, recall of 0.61, precision of 0.65 and accuracy of 0.6 across five classes. This indicates reasonably good performance, highlighting the potential of this approach.

KEYWORDS

embodied conversational agents; substance use counseling; user study; machine learning; motivational interviewing; data-sets

ACM Reference Format:
Stefan Olafsson, Byron Wallace, and Timothy Bickmore. 2020. Towards a Computational Framework for Automating Substance Use Counseling with Virtual Agents. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), Auckland, New Zealand, May 2020,* IFAAMAS, 9 pages.

1 INTRODUCTION

Over 35 million people worldwide have a substance use problem, with many struggling with opioid addiction [21]. Motivational interviewing (MI) is a face-to-face counseling method that has been shown to be particularly effective for helping individuals with substance use disorder [26]. MI aims to increase a person's motivation for change through a variety of techniques, many of which require eliciting multi-utterance, open-ended responses from clients [18].

Clients may resist resolving their substance use problem, and MI is an effective tool for counselors to decrease this resistance and move them towards willingness to take action. Client expressions that indicates this willingness are known as *change-talk* and an endorsement of the status-quo is called *sustain-talk*. Reflective listening is particularly important in MI; this entails the counselor listening intently to the client as they speak freely and then repeating, paraphrasing, or summarizing part of what was said. This technique promotes client autonomy and gives counselors the opportunity to steer the conversation in a particular direction without seeming coercive.

Given MI's effectiveness, it is unfortunate that many patients do not have access to MI counselors. And even when they do, high treatment costs may render them effectively unavailable. Automated MI-based substance use counseling with a virtual counselor thus has the potential for significant societal impact. Some MI techniques have been successfully implemented in agent-based counseling systems, for example using constrained user input [27] or spoken input using Markov decision processes [33].

However, the core techniques of MI must be used correctly for a session to be successful, namely asking open questions, affirming positive client behavior, using reflective listening to be attentive, and giving summary reflections to let the client know they're being heard [18]. These techniques require counselors to respond appropriately to richly nuanced and idiosyncratic expressions that their clients make. Additionally, counselors must recognize subtle interactional behaviors (e.g., turn-taking, engagement, avoidance, and withdrawal) and effectively manage the therapeutic agenda of the counseling session itself. Automating all of these behaviors requires models capable of reacting to complex patterns of human behavior, demanding data-driven approaches trained on corpora of successful counseling sessions.

We are developing a fully-automated virtual substance use counselor agent capable of conducting a speech-based interaction with clients, incorporating expert MI counseling techniques. This work builds on knowledge from the field of counseling psychology, past work on designing intelligent virtual agents, and uses a machine learning-based natural language processing approach to dialog systems.

In this paper, we first discuss related work, then present work towards the development of an automated counseling framework that can leverage all of MI. First, we present results from a pilot study evaluating patient acceptance of a virtual agent counselor that uses constrained user input and relatively simple dialog models. Then we describe our results from initial modeling of patient-provider dialogs to build a natural language understanding module that predicts counselor actions at every dialog turn.

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 2020, Auckland, New Zealand.

2 RELATED WORK

Our work builds on existing research into the use of virtual agents for motivational interviewing, and other forms of face-to-face counseling. We also draw from work on modeling aspects MI and patient doctor exchanges, such as automatic session transcript annotation.

2.1 Virtual Agents and Motivational Interviewing

Schulman *et al.* (2011) created a conversational agent system that had constrained user input (menu options) and used MI for long-term behavior change [27]. They developed a counseling framework that allowed the agent's dialog system to use methods like MI in its planning of therapeutic actions. In a longitudinal evaluation study, participants rated the agent higher than neutral on ratings of satisfaction, empathy, MI spirit (a measure of MI fidelity), and relational closeness, and an expert on MI rated the agent's empathy levels and MI spirit highly [1].

In a study of a similar agent for alcohol use screening and brief intervention, participants reported a therapeutic alliance significantly greater than neutral and their alcohol consumption frequency and quantity was reduced, compared to baseline [35].

Lisetti and colleagues developed and evaluated spoken dialog agent systems for conducting brief health interventions, such as motivational interviewing for people with alcohol use problems [15, 16, 33]. These systems are based on a Markov decision process framework that uses reinforcement learning on data collected from user interactions to optimize its dialog policy. They found that a system that had been optimized over user interactions achieved a higher rate of task completion, user likeability, and perceived accuracy [33]. They also developed empathic virtual agents for delivering motivational interviewing that adapt their verbal and non-verbal behavior to that of the user during counseling sessions. They found that interactions with the empathic agent compared to the non-empathic one led to more positive attitudes, higher intention to use the system again, greater perceived enjoyment, higher perceived sociability, higher perceived usefulness, greater sense of social presence, higher levels of trust, a higher rating of anthropomorphism, greater likability, higher levels of animacy, greater perceived intelligence, and greater perceived safety [16].

These systems use a limited set of techniques from MI and do not process complex unconstrained client speech in any meaningful sense, for example to generate complex reflections that act to both ground the client utterance [18] and advance the therapeutic agenda.

2.2 Modeling Motivational Interviewing Dialog using Machine Learning

Automatic annotation of MI session transcripts has the potential to allow researchers to analyze the quality of patient care at a scale not possible using manual annotation. We review prior efforts toward this end below.

Wallace *et al.* showed the feasibility of using machine learning to automatically classify utterances in transcripts of patient-provider communication. They used a conditional random field (CRF) [13] to estimate the probabilities of six relatively high-level topics, achieving an inter-rater reliability (κ) between the model and human annotators of 0.49 and an average accuracy of 0.64 [30]. They also modeled topics and speech-acts in utterances comprising patient-provider interactions jointly, which achieved better performance than a model in which topics and speech-acts were modeled independently [31]. They later extended this model to incorporate parameters representing individual doctors' speech acts, and clustered physicians on the basis of these. The induced groupings were found to correlate significantly with the scores of patient ratings of physician communication [29].

Gibson and colleagues used a Recurrent Neural Network (RNN) — specifically an Long Short-term Memory (LSTM) [9] — to: (1) classify the Motivational Interviewing Skills Code labels per utterance, and; (2) predict counselor session level empathy ratings. They used these dialog turn level behavioral acts as an encoding for a session level empathy rating. This approach outperformed training the empathy predictor without these intermediary dialog acts [7].

Pérez-Rosas *et al.* provided further evidence that RNNs are a good fit for modeling MI sessions. In this work, the authors modeled motivational interviewing sessions to automatically identify certain counselor behaviors. They annotated 277 transcribed MI sessions using a standard coding scheme, which measures counselor MI proficiency by evaluating verbal behaviors, such as reflective listening. They showed that using a feature set combining semantic and syntactic features leads to higher model performance, as compared to using bag-of-word features, and that a Gated Recurrent Unit model [4] (a particular type of RNN) achieved the highest performance for annotating counselor reflections [24].

Hasan and colleagues investigated MI counselor communication strategies from annotated transcriptions of patient-provider data. Using a model that takes the sequence of events into account, they found that there are dependencies between effective MI strategies that go beyond just the last turn of talk and provided further evidence that there are long-range dependencies in the data [8].

3 COMPUTATIONAL FRAMEWORK DEVELOPMENT

Our ultimate goal is to simulate all of MI in a natural, unconstrained, speech-based interaction between a user and a virtual counselor. However, to use a natural language interface in an automated agent introduces considerable challenges, such as: understanding natural language, deciphering pragmatics (e.g., turn-taking), and language generation for the agent's response. Our proposed framework has a dialog manager that uses modules for these specialized tasks and follows the standard architecture of a spoken dialog system (Figure 2) [34]. An input manager converts speech to text and processes the raw audio before sending it to a dialog manager. The dialog manager maintains the history of the conversation and the context, as well as any domain specific knowledge relevant to the MI counseling task. The modules component would include any task-specific models trained on patient-provider data.

Before spending resources on further framework and MI model development, however, we wanted to know if patients in treatment for substance use disorder would accept a virtual agent as a counselor. We therefore developed a prototype system and evaluated the acceptance of the interface.

Item	Anchor 1	Anchor 2	Median (IQR) - Wilcoxon
How satisfied are you with the agent?	Not at all	Very satisfied	6.5 (2.5) W=270 p<.05
How willing are you to continue working with the agent?	Not at all	Very willing	5.5 (2) W=225 p<0.05
How much do you trust the agent?	Not at all	Very much	6.5 (2) W=261 p<.05
How much do you like the agent?	Not at all	Very much	7 (1) W=306 p<.05
How repetitive was the agent?	Not at all	Very repetitive	1 (1) W=54 p<.05
How easy was it to talk to the agent?	Not at all	Very easy	7 (0.75) W=297 p<.05
How interesting was the agent?	Not at all	Very interesting	7 (2) W=261 p<.05
How would you characterize your relationship with the agent?	Complete stranger	Close friend	3.5 (2.75) n.s.
Do you feel like the agent cares about you?	Not at all	Very much	4.5 (3.25) n.s.
Do you feel like you and the agent understand one another?	Not at all	Very much	4.5 (1.75) W=216 p<.05
Was the agent honest about what she thought of you?	Not at all	Very honest	5.5 (3) W=234 p<.05
How close do you feel you and the agent are?	Not at all	Very close	2.5 (2) W=99 p<.05
How honest were you with the agent?	Not at all	Very honest	7 (0) W=324 p<.05
Would you have preferred speaking to a person about this topic?	Preferably a person	Preferably the agent	4 (1.75) n.s.

Table 1: The single item measures assessing general agent acceptance on a scale from 1-7. The last column shows whether participants' ratings were significantly different compared to a neutral rating of 4.

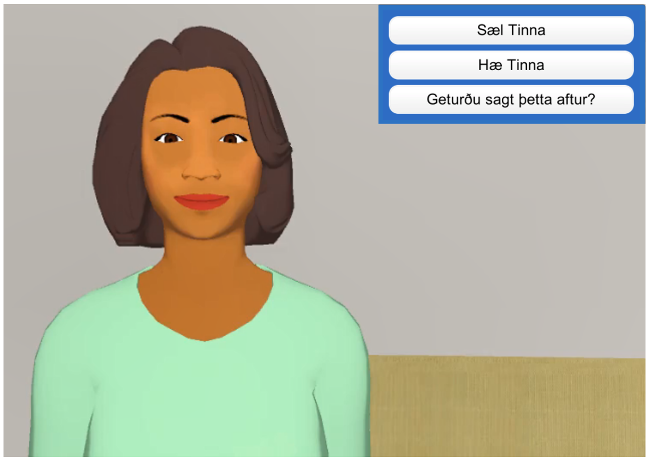

Figure 1: The opioid use disorder agent featuring a menu of user options displayed every turn.

3.1 Feasibility Pilot Study

We recruited individuals diagnosed with opioid use disorder at two outpatient clinics in two cities (approved by ethics review boards at both locations) to evaluate our prototype. Our virtual counselor is an Embodied Conversational Agent [3], an on-screen computer character that displays non-verbal conversational behavior along with its speech, providing patients with a natural and intuitive interface for face-to-face interaction. The virtual counselor was designed to support patients in medication assisted treatment for opioid use disorder; a treatment for individuals who have stopped substance use but are at risk of relapse. The agent spoke using a synthetic voice and the dialog was driven by a hierarchical task network with template-based text generation. Participants used a constrained interface for input, namely by selecting a response at every turn from a menu of options (Figure 1).

This prototype system led participants through activities shown to increase the chances of mitigating relapse, namely emotional recognition and mindfulness with deep breathing [2]. The interaction included standard MI practices, such as giving participants a chance to express whether they want to continue (or not) at crucial moments in the dialog, reflecting participant choices back to them (based on their selection from the multiple choice menu input), and using techniques like the 'readiness ruler' to measure a client's readiness to maintain abstinence.

Participants were recruited at an addiction treatment hospital in Reykjavik (Iceland) and a treatment facility in Boston (Massachusetts), with an eligibility criteria of being in medication assisted treatment for opioid use. Following consent, they filled out a demographics questionnaire, conducted a 15 minute conversation with the agent on a laptop computer, filled out a questionnaire about the interaction with the agent (Table 1), and participated in a short semi-structured interview about their experience with the agent.

A total of 23 participants successfully completed the study. Their average age was 40.22 years (SD 10.26), ranging from 23 years to 67. 22% were female, most were single, had stable housing, and had not graduated high school. The results showed that the participants were satisfied with the agent, wanted to continue working with her, trusted her, liked her, did not think that she was repetitive, felt it was easy to talk to her, and that she was interesting. Participants also felt that they and the agent understood one another, that the agent was honest about what she thought of them, that they had been honest towards the agent, and that they did not prefer speaking to a human over an agent (or vice versa) about this topic. Furthermore, participants felt that their relationship with the agent was neither close nor distant and that she neither cared too much or too little about them (Table 1).

Assessments collected from the dialog with the agent revealed that participants liked performing the deep breathing exercise with the agent (4 vs. 19, X2(1)=9.78, p<.05), and most believed that this experience will help them in their recovery (4 vs. 19, X2(1)=9.78,

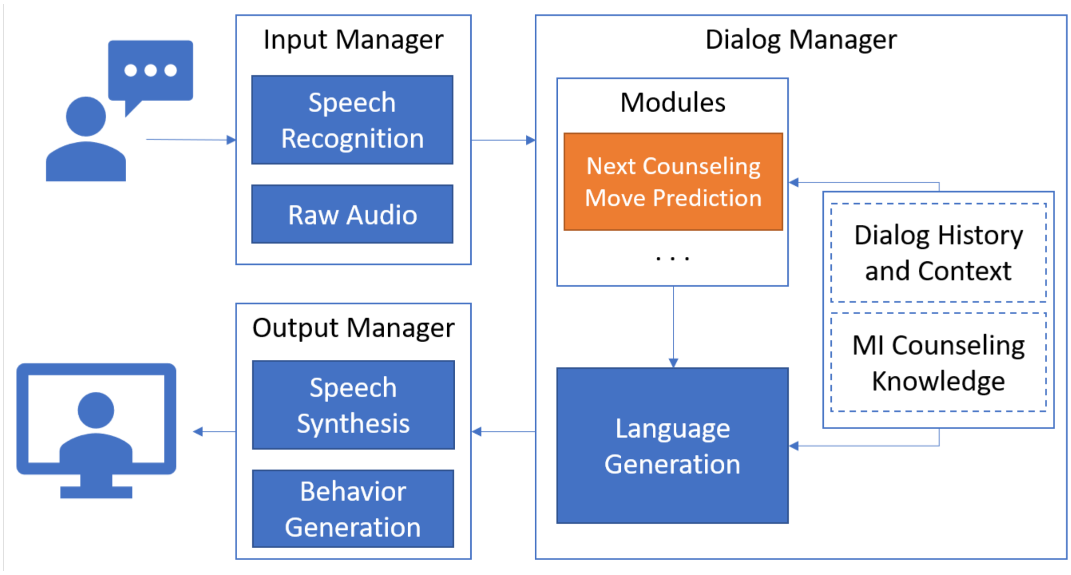

Figure 2: A diagram of the proposed computational MI counseling framework. Our current focus is predicting the next counseling move for an agent to perform (shown in orange).

p<.05). Additionally, all 23 participants were willing to self-disclose personal information about their drug use to the agent.

In the semi-structured interview, 50% of participants indicated that they would like to use a virtual agent like ours to support them in their recovery. About 35% said they would definitely use some kind of technology for support. However, 15% said they would never use any kind of technology for treatment support. Patients also suggested that the language of the interactions should be dynamically tailored to how long they've been in treatment and that the system better simulate real conversations, e.g., being allowed to speak freely as opposed to using the dialog menu options.

In summary, patients in medication assisted treatment had a generally positive reaction to a virtual agent discussing topics related to opioid use disorder therapy with them. They expressed high levels of trust in the agent and desired to work with her again. Patients were largely satisfied with the overall experience; however, measures capturing the perceived relational closeness with the agent were low. The interviews revealed that participants felt that one session was too early to talk about having any kind of relationship with the agent.

Given the positive reception of the virtual counselor prototype, we felt confident in moving forward with our framework development and took steps towards further simulation of motivational interviewing in a natural speech-based interaction.

3.2 Predicting Counselor Actions

Our first step towards a speech-based MI counseling system is predicting counselor dialog acts from text. The current focus is on creating a model that predicts the next counseling move, given the dialog context and data from annotated MI counseling session transcripts. Dialog systems typically have a defined set of user intents that the system infers given the current dialog state. Our approach directly predicts the counselor's next action at any given time, provided the dialog history and context, without explicit definition of intents or dialog state variables. This prediction can then be used, for example, to generate a response conditioned on this information and the dialog context [11].

For our experiments we used 164 annotated counseling sessions collected during brief motivational interviewing sessions about alcohol use in an emergency room setting [19]. The annotation was conducted by researchers seeking to understand the mechanisms of behavior change and were therefore coded using a variety of coding schemes, including the Generalized Behavioral Intervention Analysis System (GBIAS) and the Motivational Interviewing Skill Code (MISC) [10]. The MISC was created to assess clinician adherence to using motivational interviewing and the overall integrity of its use [20], therefore, every utterance in the data set has the MI actions of the counselor codified.

We adapted the MISC labels for our counselor move prediction task, using only the labels relevant for our purposes (Table 2). Crucially, we shifted the data set so that every set of counselor and patient utterances at time t were tagged with the counselor label at time $t + 1$; i.e., our prediction target here is the next move to take. Each session was split into multiple components, as specified by the team that annotated the original data set [10], opening up the possibility of modeling sequences of utterances at a finer level of detail than the session as a whole. Additionally, every session was split into utterances, where an utterance was defined as a stretch of text constituting a single speech-act from the GBIAS coding scheme, and each utterance annotated with a speaker label, topic, speech-act, session component, and MISC label. Table 3 shows three data samples from our data set. Each row consists of a component type, provider utterance, patient utterance, and a label denoting the next action (or move) the counselor should make.

Original MISC labels	**7 Label Experiments**	**5 labels experiment**
Open question Closed question	Question [qu] - *What does a typical week of drinking look like for you?*	Question [qu]
Simple reflection Complex reflection	Reflection [ref] - *Sounds like you're typically drinking two beers*	Reflect [ref]
Giving information Structure	Giving information [gi] - *This number shows how many drinks you had* Structure [st] - *In this first part, we'll talk a little bit*	Inform [inf]
Facilitate Filler Acknowledgement	Facilitate [fa] - *Mhmm* Acknowledgement [ack] - *Okay*	Ground [gr]
Affirm Emphasize control Support	Affirm [af] - *That was a good thing you did*	NA
All other MISC labels	Other [o]	NA
NA	NA	Shift [sh]

Table 2: The labels we used in our experiments, their corresponding original MISC label, and example utterance.

Component	**Provider Text**	**Patient Text**	**Label (next move)**
1	How does that sound?	. . . That sounds fine.	Shift
2.2	Okay		Question
2.2	So I'd like to start off with tell me about a typical week of drinking for you.	A typical week of drinking, well I don't really have a typical week of drinking. . . .	Grounding

Table 3: A snapshot of the data we used to train our models.

Focusing on this action prediction task, we conducted a series of experiments using a variety of modeling techniques and compared the results (Table 4). We split the data set into a training and validation set of 132 sessions, leaving 32 as a held-out test set. Our training set consisted of 2154 sequences (session components) and our test set had 617. A sample snapshot of the data is provided in Table 3.

The first model we considered comprised a linear support vector machine [23] over utterances represented as 'bag of words' frequency–inverse document frequency (TF-IDF) vectors. The second was a standard conditional random field (CRF) model with distributed representations of utterances extracted via Doc2Vec [14].

The third model we considered used two LSTMs to model the data. The word embedding layer was initialized via pre-training on the data set of transcripts using the Word2Vec continuous bag of words (CBOW) objective [17, 25]. Each word was embedded into a 50 dimensional space. We created separate embeddings for the patients and providers. Then, each word embedding was passed through an LSTM with hidden layer size of 64. Following each pair of patient-provider embeddings, the hidden layer of this 'word' LSTM was used as input to a second 'context' LSTM (or ConLSTM) with an input dimension of size 64 and hidden layer size of 64. We then transformed the output of the ConLSTM into a vector of dimension tagset-size and ran this through a SoftMax layer to yield probabilities over the label set; the predicted label was taken as the category associated with the highest predicted probability. Figure 3 depicts the ConLSTM architecture with an additional CRF layer.

We also experimented with the Transformer architecture [28]. Specifically, we used Bidirectional Encoder Representations from Transformers (BERT) [6], fine-tuned on our data set [32]. We performed the classification task for each utterance using the `[CLS]` embedding induced by BERT.

Another model we considered used the output from the ConLSTM as features to a CRF that then produced the final output (Figure 3). In this case, the output of the ConLSTM is then transformed into a vector of dimension tagset-size that is then used as a feature in the CRF. The CRF ultimately makes the prediction for an entire sequence jointly, which requires a 'decoding' pass.

Finally, the last model we considered consumed the output from the BERT-base model per utterance as additional features for the ConLSTM-CRF model.

All the deep neural network models were implemented using the PyTorch library [22]. Each LSTM was uni-directional, had one

Model	F1	Precision	Recall	Accuracy
7 Labels Task				
Majority label only	0.07	0.05	0.14	0.34
Linear SVC + TF-IDF	0.26	0.25	0.28	0.37
CRF + Doc2Vec	0.28	0.26	0.29	0.46
ConLSTM	0.32	0.35	0.32	0.42
BERT-base w/fine-tuning	0.37	0.34	0.37	0.47
ConLSTM-CRF	0.41	0.44	0.4	0.5
ConLSTM-CRF + BERT	0.41	0.44	0.39	0.5
5 Labels Task				
Majority label only	0.11	0.07	0.2	0.37
ConLSTM-CRF	0.62	0.65	0.61	0.59

Table 4: Results from experiments using a variety of techniques for predicting the next counselor move, given the current counselor and patient utterances and the dialog context, as well as showing a comparison to predicting the majority label only. The columns show macro-averages, treating every class equally.

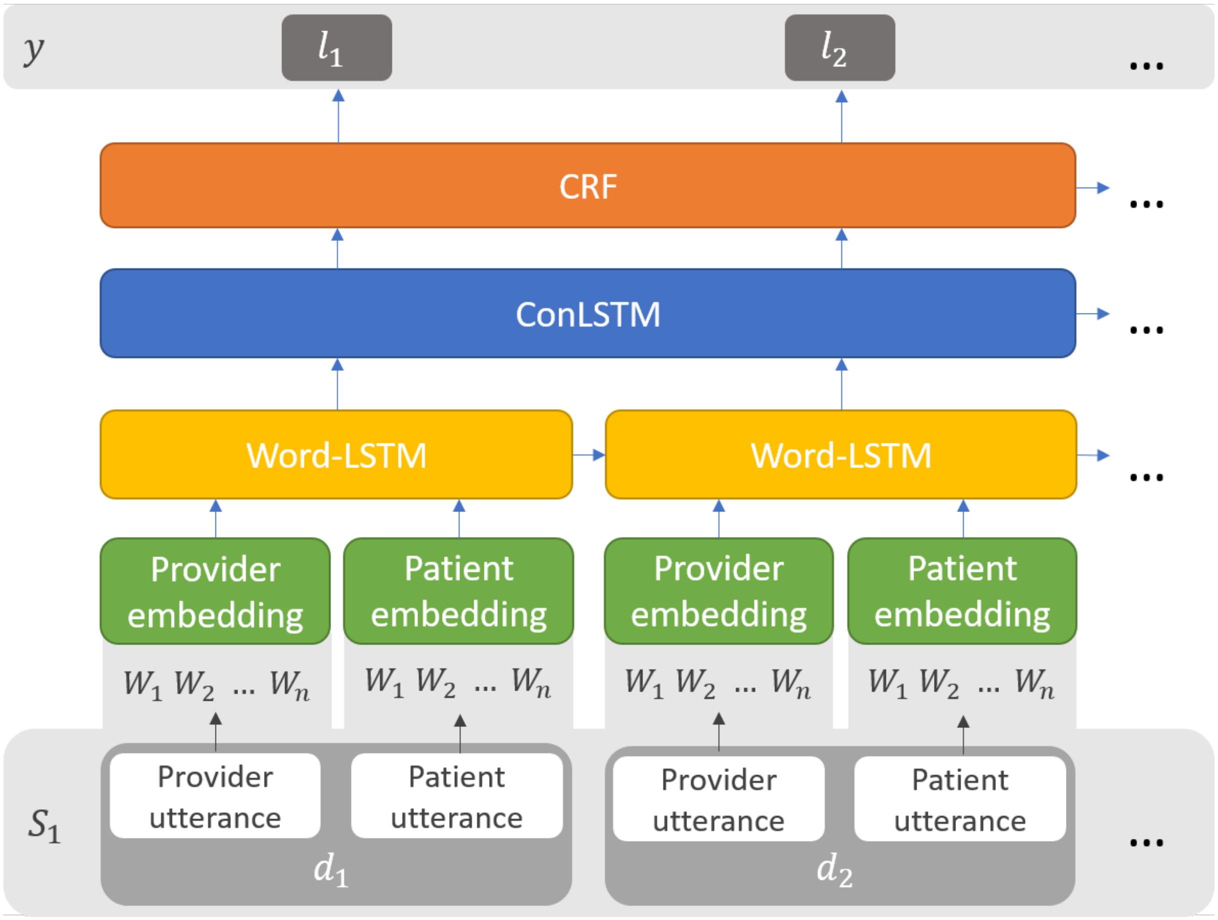

Figure 3: The ConLSTM-CRF model architecture. Each sequence (S) is a session component. Each component contains data samples (d) consisting of a label and patient-provider utterances. The model ultimately outputs a predicted label (l) for each d in the sequence.

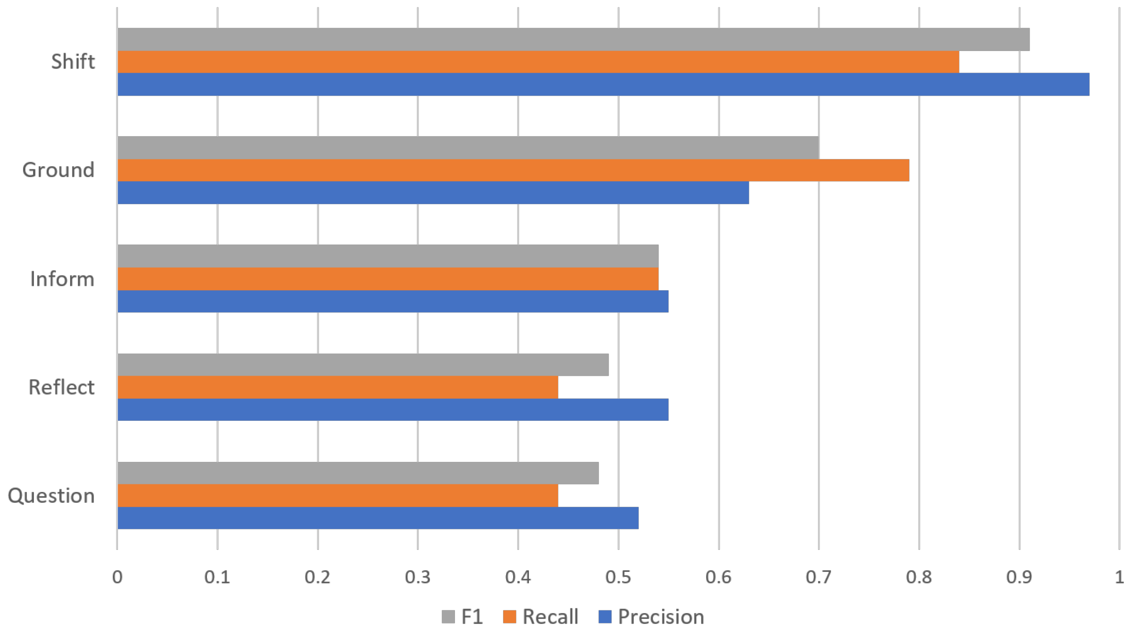

Figure 4: The macro F1, recall, and precision scores for each counseling move for our final model using the context LSTM conditional random fields approach (ConLSTM-CRF).

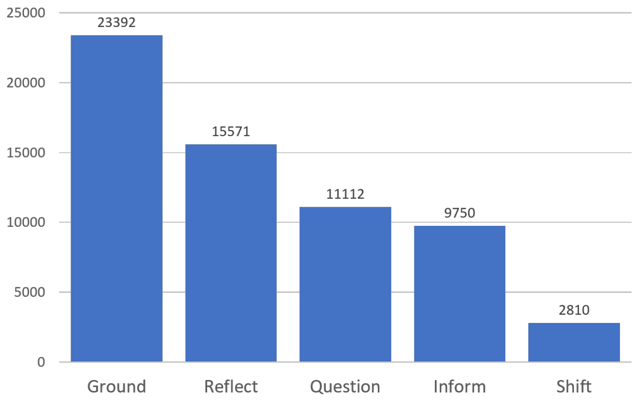

Figure 5: The frequency of each label in the training-validation data-set used for our final model.

layer, and no dropout. To fit the model we used the Adam optimization algorithm [12] using the default parameters in the PyTorch implementation: $\eta = 0.001$; $\beta = (0.9, 0.999)$; $\epsilon = 1e - 8$; $L2 = 0$.

We report macro-averaged precision, recall, and F1 scores for each model (Table 4). Precision is defined as the number of *true positives* (TP) divided by the sum of TP and *false positives*. Recall is TP divided by the sum of TP and *false negatives*. F1 is defined as

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (1)$$

Each metric has a best value of 1 and worst value of 0. These macro averages calculate metrics for each label and find their unweighted mean and do not take label imbalance into account.

We found that the ConLSTM-CRF and the ConLSTM-CRF + BERT performed the best out of these modeling approaches. Since they performed roughly the same, we chose to use the simpler model moving forward.

Following these initial experiments, we reorganized the dialog acts we were using to better reflect the types of moves we want the virtual counselor to make. First, we combined ‘acknowledgments’, ‘fillers’, and ‘facilitations’ into one *ground* move [5] and then we combined the general ‘giving information’ code and ‘session structure’ code to a new *inform* move. Lastly, we added a new move called a *shift* wherever the counselor moves to a new component of the MI session. The counts for each of the codes used in our final model are shown in Figure 5.

Using the best performing model from our previous experiments, the ConLSTM-CRF, we trained a new model. Overall, the model had an F1 of 0.62, precision of 0.65, recall of 0.61, and average accuracy of 0.59 across all five labels (Table 4). All scores presented are macro averages. Predicting the majority label only (i.e., *ground*) yielded an F1 of 0.11 and accuracy of 0.37. Looking at the results per label (Figure 4) showed that a *shift* move got the highest F1 score (F1=0.91, precision=0.99, recall=0.84), followed by *ground* (F1=0.7, precision=0.63, recall=0.79), then an *inform* move (F1=0.54, precision=0.55, recall=0.54), next a move to *reflect* (F1=0.49, precision=0.55, recall=0.44), and lastly a *question* move (F1=0.48, precision=0.52, recall=0.44).

4 CONCLUSION

We are developing a computational framework for conducting automated motivational interviewing for patients with substance use disorder using a virtual counselor. Results from a pilot study demonstrated that a virtual agent can lead sessions where some MI techniques are used and that individuals with substance use disorder find the virtual counselor acceptable.

We also reported results in using modern machine learning methods to automate additional aspects of MI. The models were trained using annotated patient-provider interaction session transcripts and we presented results from training several models that predicts a next counseling move for an agent to make, given the utterances of the provider and patient at any given dialog turn. Our best model used a custom architecture that combines a deep learning approach (LSTMs) with a sequence modeling method (CRF). The model achieved reasonable performance in terms of predicting five high-level counseling moves, indicating that this approach warrants additional research.

4.1 Limitations

Our work has several limitations. The participants in the pilot represent one set of the patients that might ultimately engage with a system such as this. Additionally, the pilot study did not have a control condition, which would have allowed us to make stronger claims about the effect of the agent.

The data we used to build and evaluate our models was collected in an emergency room setting, which may not be a context that generalizes to all use cases. Our methods are also limited in scope, the algorithms we used may not be the most fitting for the task. Therefore, we make no claims about the transferability of the particular models we trained to different domains; however, we believe similar steps can be taken to develop a counseling system for patients with different SUDs by training the models on data from the relevant domains. Further trials exploring a greater variety and combinations of models are required.

4.2 Future Work

With these limitations in mind, we continue to work towards an automated computational framework for providing MI to substance use patients. Our immediate next task is to improve on the current counselor move prediction models that we have developed.

Other near-term tasks are to add features derived from the raw audio signal to our models and study the effects of adding speech as a modality. We also aim to model other phenomena found in MI sessions, such as knowing what part of a counseling session to transition to when the agent predicts a ‘shift’ action. Finally, we plan to explore generating natural language for the virtual counselor.

REFERENCES

[1] Timothy W Bickmore, Daniel Schulman, and Candace L Sidner. 2011. A reusable framework for health counseling dialogue systems based on a behavioral medicine ontology. *Journal of biomedical informatics* 44, 2 (2011), 183–197.

[2] Sarah Bowen, Katie Witkiewitz, Seema L Clifasefi, Joel Grow, Neharika Chawla, Sharon H Hsu, Haley A Carroll, Erin Harrop, Susan E Collins, M Kathleen Lustyk, et al. 2014. Relative efficacy of mindfulness-based relapse prevention, standard relapse prevention, and treatment as usual for substance use disorders: a randomized clinical trial. *JAMA psychiatry* 71, 5 (2014), 547–556.

[3] Justine Cassell, Joseph Sullivan, Elizabeth Churchill, and Scott Prevost. 2000. *Embodied conversational agents.* MIT press.

[4] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. (2014). arXiv:cs.CL/1406.1078

[5] Herbert H Clark, Susan E Brennan, et al. 1991. Grounding in communication. *Perspectives on socially shared cognition* 13, 1991 (1991), 127–149.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] James Gibson, Dogan Can, Bo Xiao, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. 2016. A deep learning approach to modeling empathy in addiction counseling. *Commitment* 111 (2016), 21.

[8] Mehedi Hasan, April Idalski Carcone, Sylvie Naar, Susan Eggly, Gwen L Alexander, Kathryn E Brogan Hartlieb, and Alexander Kotov. 2019. Identifying effective motivational interviewing communication sequences using automated pattern analysis. *Journal of Healthcare Informatics Research* 3, 1 (2019), 86–106.

[9] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[10] Christopher W Kahler, Amy J Caswell, M Barton Laws, Justin Walthers, Molly Magill, Nadine R Mastroleo, Chanelle J Howe, Timothy Souza, Ira Wilson, Kendall Bryant, et al. 2016. Using topic coding to understand the nature of change language in a motivational intervention to reduce alcohol and sex risk behaviors in emergency department patients. *Patient education and counseling* 99, 10 (2016), 1595–1602.

[11] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858* (2019).

[12] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. (2014). arXiv:cs.LG/1412.6980

[13] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. (2001).

[14] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. 1188–1196.

[15] Christine Lisetti, Reza Amini, and Ugan Yasavur. 2015. Now all together: overview of virtual health assistants emulating face-to-face health interview experience. *KI-Künstliche Intelligenz* 29, 2 (2015), 161–172.

[16] Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. 2013. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Transactions on Management Information Systems (TMIS)* 4, 4 (2013), 19.

[17] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[18] William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change.* Guilford press.

[19] Peter M Monti, Nadine R Mastroleo, Nancy P Barnett, Suzanne M Colby, Christopher W Kahler, and Don Operario. 2016. Brief motivational intervention to reduce alcohol and HIV/sexual risk behavior in emergency department patients: A randomized controlled trial. *Journal of consulting and clinical psychology* 84, 7 (2016), 580.

[20] 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, 2 (2003), 177–184.

[21] United Nations Office on Drugs and Crime. 2019. *World Drug Report 2019.* 387 pages. https://doi.org/https://doi.org/10.18356/a4dd519a-en

[22] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic Differentiation in PyTorch. In *NIPS Autodiff Workshop*.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[24] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence Ann, Kathy J Goggin, and Delwyn Catley. 2017. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 1128–1137.

[25] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[26] Sune Rubak, Annelli Sandbæk, Torsten Lauritzen, and Bo Christensen. 2005. Motivational interviewing: a systematic review and meta-analysis. *Br J Gen Pract*

55, 513 (2005), 305–312.

[27] Daniel Schulman, Timothy Bickmore, and Candace Sidner. 2011. An intelligent conversational agent for promoting long-term health behavior change using motivational interviewing. In *2011 AAAI Spring Symposium Series*.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[29] Byron C Wallace, Issa J Dahabreh, Thomas A Trikalinos, Michael Barton Laws, Ira Wilson, and Eugene Charniak. 2014. Identifying differences in physician communication styles with a log-linear transition component model. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

[30] Byron C Wallace, M Barton Laws, Kevin Small, Ira B Wilson, and Thomas A Trikalinos. 2014. Automatically annotating topics in transcripts of patient-provider interactions via machine learning. *Medical Decision Making* 34, 4 (2014), 503–512.

[31] Byron C Wallace, Thomas A Trikalinos, M Barton Laws, Ira B Wilson, and Eugene Charniak. 2013. A generative joint, additive, sequential model of topics and speech acts in patient-doctor communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1765–1775.

[32] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771 (2019).

[33] Ugan Yasavur, Christine Lisetti, and Naphtali Rishe. 2014. Intelligent virtual agents and spoken dialog systems come together to deliver brief health interventions. *Journal on Multimodal User Interfaces, in press* 1, 1 (2014), 19.

[34] Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proc. IEEE* 101, 5 (2013), 1160–1179.

[35] S Zhou, T Bickmore, A Rubin, C Yeksigian, R Lippin-Foster, M Heilman, and SR Simon. 2017. A relational agent for alcohol misuse screening and intervention in primary care. In *CHI 2017 Workshop on Interactive Systems in Healthcare (WISH)*.