

Predicting Success of Motivational Interviews with Recurrent Neural Network and Probabilistic Model for Patient-Provider Communication Sequences

Mehedi Hasan, BS^{1*}, Alexander Kotov, PhD^{1*}, April Idalski Carcone, PhD², Ming Dong, PhD¹, Sylvie Naar, PhD²

¹Department of Computer Science, ²Department of Family Medicine and Public Health Sciences, School of Medicine, Wayne State University, Detroit, Michigan

Abstract *The problem of analyzing temporally ordered observation sequences to make predictions related to the outcome of the processes that generated these sequences arises in many domains of healthcare informatics. In this paper, we focus on patient-provider communication sequences in the context of clinical interviews and propose Recurrent Neural Network (RNN) with two baseline methods Markov chain and Hidden Markov Model (HMM), for predicting the likelihood of eliciting a particular type of patient behavioral response based on an observed sequence of patient-provider exchanges. Our method achieved 70.38%, 60.67%, 86.77% F1-score for under-sampled sequences and 77.75%, 75.20%, 83.81% F1-score for over-sampled sequences in predicting the outcome of motivational interviews with obese adolescents using Markov Chain, HMM, and RNN, respectively. The proposed method can be used to automatically identify the most effective communication strategies in motivational interviews, which significantly decreases the effort required to develop effective interventions to address many public health conditions.*

Introduction

Data in the form of temporally ordered sequences of discrete or continuous observations (e.g., symbolic sequences such as notes in patient EHR, diagnostic codes, protein or DNA sequences or continuous time series, such as ECG measurements) arise in various domains of health informatics. The order of observations in a sequence is challenging to capture using features, which makes sequence classification a more challenging task than traditional classification. In general, sequence classification methods can be divided into three categories: feature-based, distance metric-based and model-based method. Feature-based classification methods transform a sequence into a feature vector and apply a standard supervised learning algorithm, such as support vector machine¹ or decision tree². Shapelet³ and pattern^{4,5} based techniques as well as hierarchical approaches⁶ have been proposed instead of standard classifiers as well. Distance-based methods measure the similarity between sequences to determine the quality of classification. The most commonly used distance function is Euclidian distance⁷ with Dynamic Time Wrapping⁸ used for more flexible matching in time series data. The third type of sequence classification methods, first creates a probabilistic model of a sequence, such as Markov Chain (MC) and Hidden Markov Model⁹ (HMM).

However, Markov chain cannot handle long-term dependencies in sequences. Similarly, the hidden states of HMMs follow a first-order markov chain and can select only one hidden state at each time step. Therefore, Markov chain and HMMs are only capable of passing short-term dependency among sequence steps. On the other hand, Recurrent Neural Networks (RNNs) have no such limitations. RNNs are capable to capture long-term dependencies with their unlimited memory. It was shown that RNNs are very effective in sequence modeling with temporal data. Recently, two variant of RNNs is proposed: Long Short-Term Memory¹⁰ (LSTM) and Gated Recurrent Unit¹¹ (GRU), which gave superior performance in a wide range of sequence modeling problems¹²⁻¹⁴.

Sequence classification has a broad range of real word applications from genomics and health informatics to finance and anomaly detection. In genomics research, sequence classification is widely used to classify protein and text sequence data¹⁵ and detect the function of new proteins¹⁶. In health informatics, ECG measurements are considered as multi-dimensional time series and are used to classify individuals as healthy or having a heart disease¹⁷. Sequence classification is also used for anomaly detection such as abnormal access to systems¹⁸ and malware¹⁹.

In this paper, we direct our focus towards patient-provider communication sequences in the context of clinical interviews. Specifically, we focus on the transcripts of Motivational Interviews (MI) with obese adolescents and their caregivers. Childhood obesity is a serious public health concern in the United States and worldwide. Recent estimates

* Authors provided equal contribution.

indicate that approximately one third (31.8%) of US children age 2-19 years are overweight and 16.9% are obese²⁰. Adolescents who are obese likely continue to be obese in adulthood and have a greater risk of heart disease, type 2 diabetes, stroke, cancer, and osteoarthritis²¹. One approach to effective obesity intervention is Motivational Interviewing, an evidence-based counseling technique to increase intrinsic motivation and self-efficacy for health-related behavior change. The goal of MI counseling session is to encourage patients to explore their own desires, ability, reasons, need for and commitment to the targeted behavior change. These statements, referred to as “change talk” (or CT), consistently predict actual behavior change that can be sustained for as long as 34 months after an interview. However, the ability of counselors to consistently elicit this type of patient communication requires knowledge of effective communication strategies for a variety of patients, which can only be obtained through analysis of a large number of annotated interviews. Since manual annotation and examination of transcripts is a very time-consuming process, tailoring of MI interventions to particular populations can take years. Therefore, there is a need for informatics-based methods to facilitate the development of effective MI-based interventions, in general, and theoretically-grounded computational models to explore the mechanisms of MI’s efficacy, in particular.

Our research addresses this problem from multiple angles. In our previous work^{22,23}, we explored several machine learning methods for automatic annotation of clinical interview fragments with a specialized codebook containing a large number of patient and provider behavior codes²⁴. In this work, we propose Recurrent Neural Networks with two baseline probabilistic methods to identify patient-provider communication sequences that are likely to elicit the desired patient behavioral response (i.e. change talk or commitment language) and to dynamically estimate the likelihood of observing this desired response at any point during a clinical interview based on all coded previous patient-provider communication exchanges in the same interview. While there have been some previous qualitative studies of patient-provider dialog in a clinical setting²⁵, there have been no previous work on computational modeling of annotated patient-provider communication (PPC) exchanges and predicting the desired patient behavior in a context of motivational interviews.

Methods

Data collection

The experimental dataset for this work was constructed from the transcripts of 129 motivational interviews, which include a total of 50,239 segmented and annotated utterances. Each transcript consists of a counselor-adolescent and a counselor-caregiver session. Since the ultimate goal of a motivational interview is to increase the desire and ability of adolescents for a targeted behavior change, we considered only communication sequences from counselor-adolescent sessions and disregarded all communication sequences from counselor-caregiver sessions. The utterances were annotated based on MYSCOPE codebook²⁴, which are grouped into the adolescent and counselor code groups. Utterances were divided into successful and unsuccessful communication sequences. Successful communication sequences result in positive change talk and commitment language (a special class of change talk) statements by an adolescent, while unsuccessful sequences are the ones that result in negative change talk or commitment language and the sequences, in which no change talk or commitment language statements occur. Out of 5143 observed sequences, 4225 were positive and 918 were negative. For each of the probabilistic models (Markov chain and HMM), one model was trained using successful sequences and one model was trained using unsuccessful sequences. Statistics of the experimental dataset are presented in Table 1.

Table 1: Statistics of experimental dataset. Sequence length is the number of behavior codes in it.

Sequence Type	# of sequences	Ratio	Average length
Successful sequences	4225	82.15%	9.79
Unsuccessful sequences	918	17.85%	9.65

As can be seen from Table 1, our data set is imbalanced and predominately composed of successful sequences with only a small percentage (17.80%) of unsuccessful sequences. Usually, predictive accuracy misrepresents the performance of an employed algorithm for an imbalanced data. Because simply predict the majority class (successful) would provide a predictive accuracy 82.20%. Therefore, it is important to handle imbalance data properly. Many solutions

are proposed to deal with the imbalanced data sets, which can be divided into two categories: data and algorithmic levels. In this study, we utilized under and over sampling methods at data processing levels for balancing the adolescent obesity data set. Synthetic Minority Over-sampling (SMOTE) method is the most widely used method for over-sampling an imbalanced data set, in which new synthetic examples were generated from minority class²⁶. In this study, we generated synthetic examples along the borderline between minority examples and their selected nearest neighbors²⁷. On the other hand, under sampling method under samples the majority class by replacing a cluster of majority samples by the cluster centroid of a K-Means algorithm. A fragment of an adolescent session transcript is presented in Table 2. Hence, annotation column shows the successful sequence of behavior codes from top to bottom, in particular, ($SS \rightarrow OQO \rightarrow HUPO \rightarrow OQTBN \rightarrow CHT+$) is a successful behavior sequence, where counselor starts with a structured session and gets positive feedback at the end. It was noticed that same context “Yeah” and “Yes” represent different behavior codes CHT+ and HUPW, respectively. Because human brain can remember the past behaviors, and classify the current context into different behavior codes after considering their past behaviors. In this paper, we only consider those models that can capture the past events for the classification of behavior sequences.

Table 2: Fragment of the annotated transcript of a dialogue between a counselor and an adolescent.

Annotation	Behavior	Speaker	Text
SS	Structure Session	Counselor	Okay. Can I meet with Xxxx alone for a few minutes?
OQO	Open-ended question, other	Counselor	So, Xxxx, how you doing?
HUPO	High uptake, other	Adolescent	Fine
OQTBN	Open-ended question, target behavior neutral	Counselor	That’s good. So, tell me how do you feel about your weight?
CHT+	Change Talk positive	Adolescent	It’s not the best.
CQECHT+	Closed question, elicit change talk positive	Counselor	It’s not the best?
CHT+	Change Talk positive	Adolescent	Yeah
CQTBN	Closed question, target behavior neutral	Counselor	Okay, so have you tried to lose weight before?
HUPW	High uptake, weight	Adolescent	Yes

Prediction method

Generally, a sequence is a temporally ordered set of events. In this study, an event is a behavior code that also has a symbolic representation, such as $LUP+$ (low uptake, positive), $OQECHT+$ (open-ended question, elicit change talk positive), etc. Given a sequence of behavior codes $S_i = \{c_1, c_2, \dots, c_n\}$ representing patient-provider communication exchanges during some part of a motivational interview, the task of predicting interview success can be viewed as sequence classification. Given a set of class labels $L = \{l_1, l_2, \dots, l_m\}$ (in our case, the labels are “successful” and “unsuccessful” motivational interview), a sequence classifier C learns a function $S_i \rightarrow l_i, l_i \in L$ that maps a sequence S_i into a class label $l_i \in L$.

Our proposed baseline prediction method consists of two steps. In the first step, we model successful and unsuccessful patient-provider interactions using first- and second-order Markov Chain and Hidden Markov Model, which are popular probabilistic models for discrete observation sequences with finite vocabulary. In the second step, we classified each test sequence based on the maximum likelihood of generating that sequence from each model. Although HMM was originally developed for speech recognition⁹, it is one of the most widely used methods for sequence modeling^{28–30}. However, the latest advances in deep learning technologies show that deep learning, in particular, RNNs provide better results than conventional machine learning methods for the task of sequence classification. Therefore, we employed two special variants of RNN models in our experimnts: long short-term memory (LSTM) and gated recurrent unit (GRU).

Markov Chain (MC) is a stochastic model for randomly changing systems, which assumes that the next state of

a system only depends on its current state and not on its past states (Markov property). Generally, this assumption enables reasoning and computation with the models that would otherwise be intractable. For the sequential analysis, we built two Markov models M and \bar{M} summarizing provider strategies and patient responses in case of successful (M) and unsuccessful (\bar{M}) motivational interviews. A Markov model M can be represented as a weighted directed graph $G = (V, E, p)$, in which:

- $V = \{CML+, CHT+, CHT-, AMB-, BLT, LUP+, LUP-, HUPW, \dots\}$ is a set of vertices, consisting of adolescent and counselor MI behavior codes;
- $E \subseteq V \times V$ is a set of edges corresponding to possible transitions from one MI behavior code to the other in a sequence;
- $p_M : E \rightarrow [0..1]$ is a function that assigns probability $p(c_i|c_j)$ to an edge between the MI behavior codes c_i and c_j based on maximum likelihood estimator:

$$P_M(c_j|c_i) = \frac{n_{c_i, c_j}}{n_{c_i}} \quad (1)$$

where n_{c_i, c_j} and n_{c_i} are the number of times a transition between the MI behavior codes c_i and c_j and the code c_i has been observed in the training data. Given a Markov model M (such that $S \subseteq V$), the probability that a sequence of MI behavior codes $S = \{C_1, \dots, C_N\}$ has been generated from a Markov model M is:

$$P_M(S) = \prod_{i=2}^N p_M(c_i|c_1, \dots, c_{i-1}) = \prod_{i=2}^N p_M(c_i|c_{i-1}) \quad (2)$$

In the second step, we quantify the likelihood of success of a given motivational interview at a certain time point given a sequence of MI behavior codes S observed prior to that point using the following formula:

$$p(S \rightarrow CML+) = \log \left(\frac{P_M(S)}{P_{\bar{M}}(S)} \right) = \sum_{i=2}^N \log p_M(c_i|c_{i-1}) - \sum_{i=2}^N \log p_{\bar{M}}(c_i|c_{i-1}) \quad (3)$$

If $p(S \rightarrow CML+) > 0$, the interview is predicted to result in positive change talk or commitment language.

The Markov model we have discussed so far is referred to as the first-order MC, since it only considers immediately preceding behavior code when computing the state transition probabilities. A generalization of the first-order MC is the k^{th} order MC, in which the transition probability to a particular state (symbol) x_i is computed by looking at the k preceding states (symbols). Thus, k^{th} order Markov chain will have N^k states each associated with a sequence of k symbols. In our experiment, we also used second order Markov model. Hence, our model has N^2 states with transition probability matrix of size $N^2 \times N$.

Hidden Markov Model (HMM) is a powerful tool for statistical modeling of processes varying in time. HMMs are widely used for sequence analysis because of their ability to incorporate dependencies among elements in sequence. HMM can be considered as a doubly embedded stochastic process with a process that is not observable (hidden process) and can only be observed through another stochastic process (observable process) that produces a sequence of observations. An HMM can be fully specified by the following quintuple: $\lambda = (N, M, A, B, \pi)$

- N is a number of hidden states in the model
- M is a number of distinct observations symbols per state, i.e. the discrete vocabulary size
- A is an $N \times N$ state transition probability distribution matrix $A = \{a_{ij}\}$
- B is an $N \times M$ matrix $B = \{b_j(k)\}$ with observation symbol probability distribution for each state
- π is the initial state distribution vector $\pi = \{\pi_i\}$

We exclude the structure parameters M and N to designate HMM using a compact notation: $\lambda = (A, B, \pi)$. The key difference between HMM and MC is that HMM requires specifying the number of hidden states as a model parameter and then the model deduces a sequence of hidden states that best explains the observations along with state transition probabilities and distributions of observation symbols for each state. The Baum-Welch algorithm was used to estimate the parameters of HMMs for successful and unsuccessful interviews using the corresponding training set, while the Viterbi algorithm was used to determine the most likely sequence of hidden states for a given sequence of observations. After assignment of hidden states, the log-likelihood of successful outcome can be estimated using Eq. 3.

MYSCOPE code embedding: We utilize word embedding in place of one-hot vectors for the representation of behavior code. Because one-hot vectors are high-dimensional and sparse. On the other hand, word embeddings have the ability to represent semantically similar words with similar vectors in a lower dimensional space by retaining their internal relations. Figure 1 illustrates the myscope code embedding visualize with t-SNE, a sophisticated technique for visualizing high-dimensional data. It can be seen that similar behavior codes are placed close to each other in high dimensional vector space such as OQECHT+, OQECML+, AF, ADV+, RCML+S, etc are nearest neighbors of CQECML+, which make sense in real world practice.

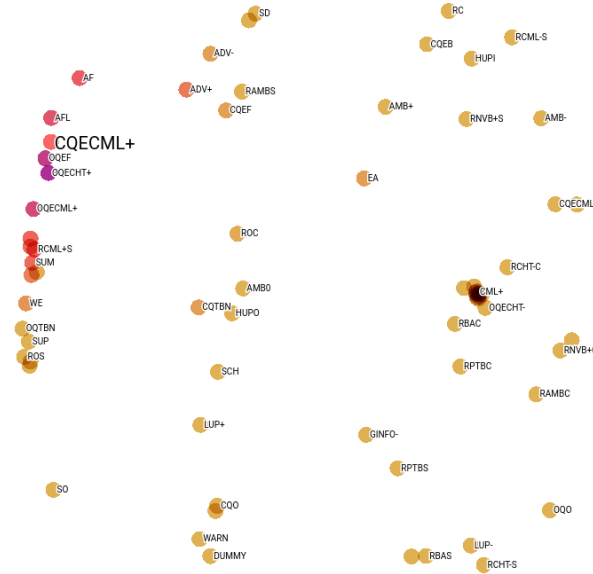


Figure 1: T-SNE diagram of behavior code drawn after 4000 steps with perplexity 30.

Recurrent Neural Network (RNN) is a type of Neural Network, which can send feedback from the current hidden state to the hidden state of the next time step. In this way, RNN can capture long-term dependencies for predicting the future events, which is the main advantage of RNN over feedforward Neural Network, MC, and HMM. The capability of remembering past event is very useful in motivational interviews where a behavior observed at some points in the interview is very informative for the future behaviors that will be observed. However, it was observed that RNN fails to capture long-term dependencies due to vanishing gradient problem³¹. In order to mitigate this problem, Hochreiter et al.³² proposed a special kind of RNN called Long Short Term Memory networks or simply LSTM. Later, the model was enhanced by including forget gates and the tanh activation function¹⁰. There are several variants of LSTM model, where a dramatic variation on the LSTM is the Gated Recurrent Unit³³ or simply GRU. GRUs are simpler than LSTM units and experimentally outperform than other models in many cases. We reiterate the mathematical formulation of GRU defined by Chung et al.¹¹ as follows (Eq. 4 to 7):

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (4)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (5)$$

$$\tilde{h}_t = \tanh(W_h x_t + r_t \odot U_h h_{t-1} + b_h) \quad (6)$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \quad (7)$$

$$total\ loss = \alpha \cdot \frac{1}{T} \sum_{t=1}^T loss(\bar{y}^{(t)}, y^{(t)}) + (1 - \alpha) \cdot loss(\bar{y}^{(T)}, y^{(T)}) \quad (8)$$

The architecture of the proposed Recurrent Neural Network is shown in Figure 2. We used one hidden layer of 20

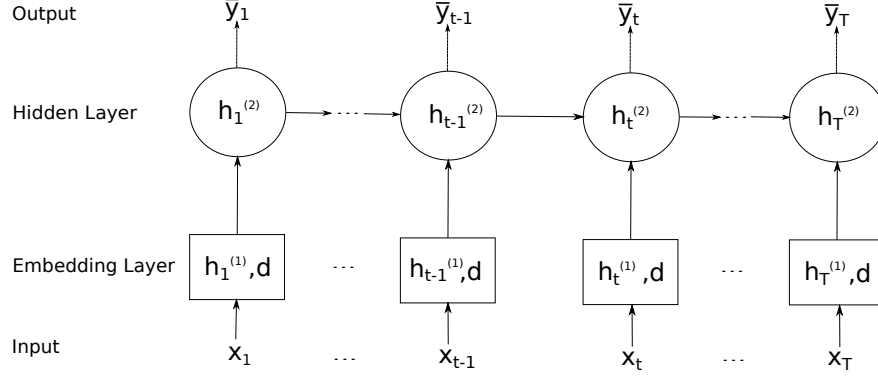


Figure 2: Proposed RNN model with target replication (TR).

LSTM or GRU nodes after embedding the given myscope codes from the input sequence. Then, we applied a softmax layer at each time step for predicting the target. This target replication (TR) strategy was used in several studies^{13,14} after introduced by Lee et al.³⁴ in 2015 for convolutional neural networks. As it can be seen from equation (8), our loss is the combination of final loss and the mean of the losses over all sequence steps, where T is the total number of sequence step, \bar{y} is the output at step t , and $\alpha \in [0, 1]$ is a hyperparameter that determines the relative importance of intermediate outputs. Our model contains several hyperparameters such as embedding dimension, number of hidden units, learning rate, batch size, etc., which were determined by the validation set. We implemented our models in Tensorflow with adam optimizer and early stopping based on the validation loss. Experimentally, we observed that our model converges to optimal after 100 epoch and have no effect of dropout and regularization. The code for all models is publicly available at <https://github.com/teanalab/myscope-sequential-analysis>.

Evaluation metrics

Performance of the proposed method was evaluated in terms of precision, recall, and F-measure using 10 folds cross-validation and weighted macro-averaging of these metrics over the folds. However, LSTM and GRU are trained on 70% of the data and validated on 20%. The remaining 10% of the data is used as a test set for reporting the performance of the model. A true positive (TP) was counted when the method correctly classified a sequence into its actual class; a false positive (FP) was counted for a class when the method incorrectly classified a sequence into that class; a false negative (FN) for an actual class of sequence was counted when the method incorrectly classified the sequence into other class. The Precision of a class was defined as the ratio of the numbers of correctly classified sequences and all sequences identified as belonging to that particular class by the classifier (i.e. Precision = TP / (TP + FP)). Recall of a class was defined as the ratio between the numbers of correctly classified sequences and all sequences of that particular class in the gold standard (i.e. Recall = TP / (TP + FN)). F-measure is computed as the harmonic mean of precision and recall (i.e. $2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$).

Results

Experimental evaluation of the proposed method is conducted on both under- and over-sampled sequences.

Table 3: Performance of MC, HMM, and RNN for predicting the success of under- and over-sampled patient-provider communication sequences. The highest value for each performance metric is highlighted in bold.

Method	Under Sampling			Over Sampling		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Markov Chain 1 st Order	0.7060	0.7044	0.7038	0.7932	0.7799	0.7775
Markov Chain 2 nd Order	0.6395	0.6385	0.6379	0.7111	0.7029	0.7000
Hidden Markov Model	0.6244	0.6143	0.6067	0.7775	0.7567	0.7520
LSTM RNN	0.8672	0.8626	0.8622	0.8411	0.8372	0.8368
LSTM RNN - TR	0.8733	0.8681	0.8677	0.8424	0.8385	0.8381
GRU RNN	0.8674	0.8648	0.8646	0.8379	0.8342	0.8337
GRU RNN - TR	0.8705	0.8676	0.8673	0.8412	0.8377	0.8373

Predictive performance for under-sampled PPC code sequences

We applied small learning rate 0.00005 and early stopping strategy with 150 epoch for training deep learning models when under-sampled data is used. Predictive performance summary of the proposed methods on under-sampled sequences is presented in Table 3. Five major conclusions can be drawn from the results in Table 3. First, recurrent neural networks outperform over probabilistic models and achieved 18.50%-39.88% higher F1-score. Second, GRU with target replication has the best performance over all other RNN methods, and achieved accuracy 0.8858, F1-score 0.8853 with precision 0.8917 and recall 0.8858. Third, target replication strategy improves the performance of GRU and LSTM, while conventional GRU shows better performance than LSTM with target replication. Fourth, among probabilistic models, the MC-based method generally outperforms HMM across all metrics for under-sampled sequences. Fifth, it follows that second-order MC have lower accuracy, precision, recall, and F-measure than first-order MC. In particular, accuracy, precision, recall and F-measure decrease by 15.80%, 15.92%, 15.80% and 15.89%, when going from first- to second-order models. At last, accuracy and recall illustrate exactly the same values for all models.

Predictive performance for over-sampled PPC code sequences

For over-sampled data, RNN models are trained for 100 epoch with early stopping strategy. Table 3 summarizes predictive performance of the proposed methods on over-sampled sequences. These results indicate that HMM had better performance than second-order MC, achieving 5.74% higher F1-score, while HMM still has 2.41% lower F1-score than first-order Markov chain. Similar to the case of under-sampled sequences, target replication improves the performance of RNN models, and GRU with target replication has the highest F1-score of 0.8716 among all models. However, the predictive performance of recurrent neural networks decreases in over-sampled sequences whereas the performance of probabilistic models decreases. We also observed that all models have the largest value in precision compared to other performance metrics.

Most likely communication sequences

Table 4 provides examples of typical PPC sequences that frequently appear in successful and unsuccessful motivational interviews. It can be seen that most successful sequence start with a summary of the discussion or structured statement. After that, if adolescents express positive change talk, the counselor immediately reflects on that to reinforce adolescent's intrinsic motivation about behavior change. On the other hand, providing additional information can lead to negative change talk, even in the cases when adolescents were showing a positive tendency in their previous communication. This observation can be explained by adolescents quickly losing focus when provided with general information that undermines their motivation. Analyzing such cases will allow the counselors to determine the general information that can be provided during the interviews.

Discussion

By analyzing the experimental results of different communication sequence outcome prediction schemes proposed in this paper, we arrived at the following conclusions. First, using higher-order models results in better prediction

Table 4: Most likely communication sequences in successful and unsuccessful motivational interviews.

Type	Most likely communication sequences
successful	GINFO+: General information, positive → LUP+: Low uptake, positive → OQTBN: Open-ended question, target behavior neutral
successful	SS: Structure session → GINFO+: General information, positive → CQECHT+: Closed-ended question, elicit change talk positive
successful	SO: Statement, other → LUP+: Low uptake, positive → AF: Affirm → HUPW: High uptake, weight → OQECML+: Open-ended question, elicit commitment language positive.
unsuccessful	ADV+: Advise, positive → AMB-: Ambivalence negative → OQECHT-: Open-ended question, elicit change talk negative
unsuccessful	CQECHT+: Open-ended question, elicit change talk positive → RCHT-S: Reflect, change talk negative → OQECHT-: Open-ended question, elicit change talk negative
unsuccessful	SUP: Support → AF: Affirm → CQTBN: Closed-ended question, target behavior neutral → OQECHT-: Open-ended question, elicit change talk negative → AMB-: Ambivalence negative

accuracy compared to lower-order models, since higher-order models utilize larger context and can better capture different nuances of patient-provider communication. On the other hand, the number of states in higher-order Markov models grows exponentially with their order. Therefore, accurate estimation of transition probabilities requires much larger training set. Using smaller datasets will result in a sparsity problem, when many transitions are either not observed in the training set at all or observed only a few times, leading to missing or potentially inaccurate probability estimates. Obtaining large training sets cannot be easily accomplished in many domains, including motivational interviewing. In this study, we found out that using second-order Markov models is a reasonable trade-off between efficiency and accuracy.

Second, the overall predictive performance of HMM is substantially better than MC for both types of sequences. In particular, HMM-based method achieves near-human accuracy for predicting the success of motivational interviews. This indicates that hidden states in HMM are able to capture the structure of discourse in motivational interviews by grouping together the codes that correspond to different cognitive states of adolescents during motivational interviews, which reflect the overall progression of the interviews. This allows to reduce the dimensionality of codes in PPC sequences and consequently improve both precision and recall of the prediction method.

Third, converting natural PPC communication sequences to alternating ones results in better performance for all configurations of the proposed method. This indicates that alternating sequences emphasize the dependencies between the pairs of patient and provider codes, which results in more accurate estimates of the corresponding conditional probabilities. Better estimates of parameters in Markov models of successful and unsuccessful interviews are propagated to the next step, where they are aggregated into predictions for the entire sequence. This allows to achieve a dramatic improvement in the prediction accuracy of the method.

Fourth, the proposed method can be used to identify the most effective communication strategies at eliciting a particular type of behavioral response. Awareness of these strategies by researchers can significantly decrease the time and effort required to develop effective interventions to address many public health conditions, such as childhood obesity, and tailor these interventions to particular patient cohorts. Awareness of these strategies by the counselors can lead to greater success rate of motivational interviews.

Conclusion

In this paper, we proposed Recurrent Neural Networks with two baseline methods Markov Chain and Hidden Markov Model for predicting the success of motivational interviews. We found out that individual patient-provider communication exchanges are highly indicative of the overall progression and future trajectory of clinical interviews and can be used to predict their overall success. Our proposed method can facilitate motivational interviewing researchers in establishing causal relationships between different communication strategies and the desired behavioral outcomes during the interviews without resource intensive manual qualitative analysis of interview transcripts. Our proposed method can also help to identify specific patterns that are common to successful and unsuccessful motivational inter-

views, which can also directly inform clinical practice. This work also has broad implications for qualitative public health research by providing a formal theoretically-grounded computational mechanism to facilitate the development of effective behavioral interventions.

Acknowledgments

This study was supported by an R21 grant DK108071 from the NIH. We would like to thank the student assistants in the Department of Family Medicine and Public Health Sciences at Wayne State University School of Medicine for their help with transcribing the recordings of motivational interviews.

References

- [1] Leslie C, Kuang R. Fast string kernels using inexact matching for protein sequences. *Journal of Machine Learning Research*. 2004;5(Nov):1435–1455.
- [2] Chuzhanova NA, Jones AJ, Margetts S. Feature selection for genetic sequence classification. *Bioinformatics*. 1998;14(2):139–143.
- [3] Ye L, Keogh E. Time series shapelets: a new primitive for data mining. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM; 2009. p. 947–956.
- [4] Kudenko D, Hirsh H. Feature generation for sequence categorization. In: *AAAI/IAAI*; 1998. p. 733–738.
- [5] Lesh N, Zaki MJ, Ogihara M. Mining features for sequence classification. In: *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM; 1999. p. 342–346.
- [6] Nallam K, et al. An Effective Implementation of Hierarchical Approach For Sequence Classification. *IJACTA*. 2016;3(2):143–146.
- [7] Keogh E, Kasetty S. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and knowledge discovery*. 2003;7(4):349–371.
- [8] Keogh EJ, Pazzani MJ. Scaling up dynamic time warping for datamining applications. In: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM; 2000. p. 285–289.
- [9] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*. 1989;77(2):257–286.
- [10] Graves A, Mohamed Ar, Hinton G. Speech recognition with deep recurrent neural networks. In: *Acoustics, speech and signal processing (icassp), 2013 IEEE international conference on*. IEEE; 2013. p. 6645–6649.
- [11] Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*. 2014;.
- [12] Nion T, Menasri F, Louradour J, Sibade C, Retornaz T, Métaireau PY, et al. Handwritten information extraction from historical census documents. In: *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE; 2013. p. 822–826.
- [13] Lipton ZC, Kale DC, Elkan C, Wetzell R. Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677*. 2015;.
- [14] Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor ai: Predicting clinical events via recurrent neural networks. In: *Machine Learning for Healthcare Conference*; 2016. p. 301–318.
- [15] Yakhnenko O, Silvescu A, Honavar V. Discriminatively trained markov model for sequence classification. In: *Data Mining, Fifth IEEE International Conference on*. IEEE; 2005. p. 8–pp.

- [16] Deshpande M, Karypis G. Evaluation of techniques for classifying biological sequences. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer; 2002. p. 417–431.
- [17] Wei L, Keogh E. Semi-supervised time series classification. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2006. p. 748–753.
- [18] Lane T, Brodley CE. Temporal sequence learning and data reduction for anomaly detection. *ACM Transactions on Information and System Security (TISSEC)*. 1999;2(3):295–331.
- [19] Drew J, Hahsler M, Moore T. Polymorphic malware detection using sequence classification methods and ensembles. *EURASIP Journal on Information Security*. 2017;2017(1):2.
- [20] Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of obesity and trends in body mass index among US children and adolescents, 1999-2010. *Jama*. 2012;307(5):483–490.
- [21] General US. Surgeon Generals vision for a healthy and fit nation. Washington, DC: HHS. 2010;.
- [22] Kotov A, Hasan M, Carcone A, Dong M, Naar-King S, BroganHartlieb K. Interpretable probabilistic latent variable models for automatic annotation of clinical text. In: *AMIA Annual Symposium Proceedings*. vol. 2015. American Medical Informatics Association; 2015. p. 785.
- [23] Hasan M, Kotov A, Carcone AI, Dong M, Naar S, Hartlieb KB. A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. *Journal of biomedical informatics*. 2016;62:21–31.
- [24] Carcone AI, Naar-King S, Brogan K, Albrecht T, Barton E, Foster T, et al. Provider communication behaviors that predict motivation to change in black adolescents with obesity. *Journal of developmental and behavioral pediatrics: JDBP*. 2013;34(8):599.
- [25] Eide H, Quera V, Graugaard P, Finset A. Physician–patient dialogue surrounding patients expression of concern: applying sequence analysis to RIAS. *Social Science & Medicine*. 2004;59(1):145–155.
- [26] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*. 2002;16:321–357.
- [27] Nguyen HM, Cooper EW, Kamei K. Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*. 2011;3(1):4–21.
- [28] Mutsam N, Pernkopf F. Maximum margin hidden Markov models for sequence classification. *Pattern Recognition Letters*. 2016;77:14–20.
- [29] Won KJ, Prügél-Bennett A, Krogh A. Training HMM structure with genetic algorithm for biological sequence analysis. *Bioinformatics*. 2004;20(18):3613–3619.
- [30] Chai W, Vercoe B. Folk music classification using hidden Markov models. In: *Proceedings of International Conference on Artificial Intelligence*. vol. 6. sn; 2001. .
- [31] Bengio Y, Frasconi P, Simard P. The problem of learning long-term dependencies in recurrent networks. In: *Neural Networks, 1993., IEEE International Conference on*. IEEE; 1993. p. 1183–1188.
- [32] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural computation*. 1997;9(8):1735–1780.
- [33] Cho K, Van Merriënboer B, Bahdanau D, Bengio Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:14091259*. 2014;.
- [34] Lee CY, Xie S, Gallagher P, Zhang Z, Tu Z. Deeply-supervised nets. In: *Artificial Intelligence and Statistics*; 2015. p. 562–570.