Designing an effective intervention of motivational interview using sequence classification methods

Mehedi Hasan, BS[1], Alexander Kotov, PhD[1], April Carcone, PhD[2], Sylvie Naar-King, PhD[2]
[1]Department of Computer Science, [2]Pediatric Prevention Research Center, Wayne State University, Detroit, Michigan

Abstract *Coming soon!*

Introduction

In the recent times, more and more data from various domains being produced in the form of event sequences. Thus, sequence classification has become an important task in sequence mining. Sequences can contain discrete (e.g., symbolic sequences such as text, proteins, or DNA) or continuous events (e..g, time series such as ECGs or stocks), depending on the event types. Usually, sequences don't have explicit features and suffer with high dimension of feature space. Even the sequential nature of features is too difficult to capture, which make sequence classification, a more challenging task than feature-based classifications.

Sequence classification has a broad range of real word applications such as information retrieval, genomic analysis, health informatics, finance and abnormal detection.

Annotation of clinical interview transcripts to distinguish different behavior types is an important part of clinical research aimed at designing effective interventions for many conditions and disorders. In this paper, we focus on the transcripts of Motivational Interviews with obese adolescents (teens) and their caregivers. Childhood obesity is a serious public health concern in the United States and worldwide. Recent estimates indicate that approximately one third (31.8%) of US children age 2-19 years are overweight and 16.9% are obese 01. Adolescents who are obese are likely continue to be obese in adulthood and have a greater risk of heart disease, type 2 diabetes, stroke, cancer, and osteoarthritis 02. Therefore, there is a need for informatics-based methods to facilitate development of effective interventions for childhood obesity. One approach to designing intervention is Motivational Interviewing (MI), an evidence-based counseling technique to increase intrinsic motivation and self-efficacy for health-related behavior change 03,04. The goal of a 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 05 that can be sustained for as long as 34 months after an interview 06.

Automatic annotation of patient utterances in clinical communication is a challenging task, since patients usually come from a variety of cultural and educational backgrounds and their language use can be quite different 09. This problem is exacerbated when the interviews are conducted with children and adolescents due to their tendency to use incomplete sentences and frequently change subjects.

Previous quantitative studies of clinical conversation have resulted in creation of Generalized Medical Interaction Analysis System (GMIAS) 10, which uses a codebook with generic hierarchical categories. The small-size codebook in Comprehensive Analysis of the Structure of Encounters System (CASES) 11 was designed to annotate several meta-discursive aspects of medical interviews, such as assigning "ownership" of topics and partitioning them into distinct segments (speech acts). It was also shown that the fragments of transcripts of routine outpatient visits consisting of several speech acts coded using GMIAS and CASES can be annotated as "information giving" and "requesting information" 12. Other related previous studies focused on

In this project, we used a probabilistic model, in particular, first order markov model for the analysis of sequential data to determine provider-patient communication sequences that are likely to translate into change talk and commitment language.

Methods

Data collection

The MYSCOPE 07 codebook contains a total of 115 different codes that are grouped into the youth, caregiver, and counselor code groups. The experimental datasets for this work were constructed based on the transcripts of 37 motivational interview sessions, which include a total of 11,353 segmented and annotated utterances. These utterances have been further partitioned into two subsets based on the structure of motivational interview sessions: one dataset that includes all utterances from the adolescent sessions (6,579 samples) and the other dataset that includes all utterances from the caregiver sessions (4,774 samples). A fragment of an adolescent session transcript is presented in Table 1.

Table 1: Fragment of the annotated transcript of a dialogue between a counselor and an adolescent.

Annotation	Description	Speaker	Text
331	Open-ended question, elicit change talk positive	Counselor	do you feel like making healthier choices for your snacks and your meals is something you would be able to do ? mm-hmm meaning is that food available for you ?
117	Low Uptake, positive	Adolescent	Yes
301	Structure Session	Counselor	okay and thats an important thing for us to think about cause i would not want to help you come up with a plan that you would not be able to do without somebody else help so the last part of your plan is how somebody could be supportive to you meaning how they can help you be successful and so we should choose somebody who you feel like is around often enough
112	Change Talk positive	Adolescent	my um aunt
301	Structure Session	Counselor	okay so lets stick something my aunt can do
112	Change Talk positive	Adolescent	she could when i am doing when i am eating something that i should i could not be eating but so i can choose something healthy she could tell me not to eat it
309	Affirm, low	Counselor	okay that sounds like a really great suggestion

To create the gold standard for our second experiment, we partitioned the adolescent transcripts into sequence of successful and unsuccessful codes which are ended with either negative or positive change talk and commitment language. Table **??** depicts the distribution of PPC code sequences over the adolescent dataset.

System overview

The pipeline consists of two stages: training and testing. Prior to the training stage, we preprocess the collected clinical interview transcripts by performing stemming, punctuation removal, word segmentation and tokenization. Features are then extracted from the preprocessed data. During this stage, previous label and LIWC features are used in conjunction with the lexical features to create the feature vectors. After that, classifiers are trained on the feature vectors extracted from the training samples and their associated annotations. In the testing stage, after creation of feature vectors, the previously trained classifiers predict the label of each utterance in the testing sample. Finally, performance of different classifiers is evaluated by calculating standard metrics such as precision, recall, F-score (F1), kappa measure and accuracy. Specifically, we evaluated the performance of the following state-of-the-art supervised machine learning methods.

Markov Model: In probability theory, a markov model 45 is a stochastic model used to model randomly changing systems where it is assumed that future states depend only on the present state and not on the sequence of events that preceded it (that is, it assumes the markov property). Generally, this assumption enables reasoning and computation with the model that would otherwise be intractable. For the sequential analysis, we built two markov models M and $\overline{M}$ describing provider strategies and patient responses in case of successful (M) and unsuccessful ($\overline{M}$) motivational interviews. A markov model M can be represented as a weighted directed graph $G = (V, E, p)$, in which:

- $V = \{CML+, CHT+, CHT-, T-AMB, CCT, BLT, LUP+, LUP-, HUP-W, ...\}$ is a set of vertices, consisting of possible youth and counselor MI behavior codes;
- $E \subseteq V \times V$ is a set of edges corresponding to posssible 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:

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

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 have been observed in the training data. Given a markov model M (such that $S \subseteq V$), a sequence of MI behavior codes $S = \{C_1, ..., C_N\}$ have been generated from a markov model M is:

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

Success of a given motivational interview, given a sequence of MI behavior codes corresponding to it, can be predicted based on the following formula:

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

if $p(S \rightarrow CT) > \delta$, where δ is experimentally defined threshold, the interview transcript (or a portion of it) is predicted to result in CT. We experimentally determined the value of δ to 0.3 which provides acceptable accuracy with minimum false positive rate. Figure1 illustrates the characteristics of proposed model by varying the value of δ.

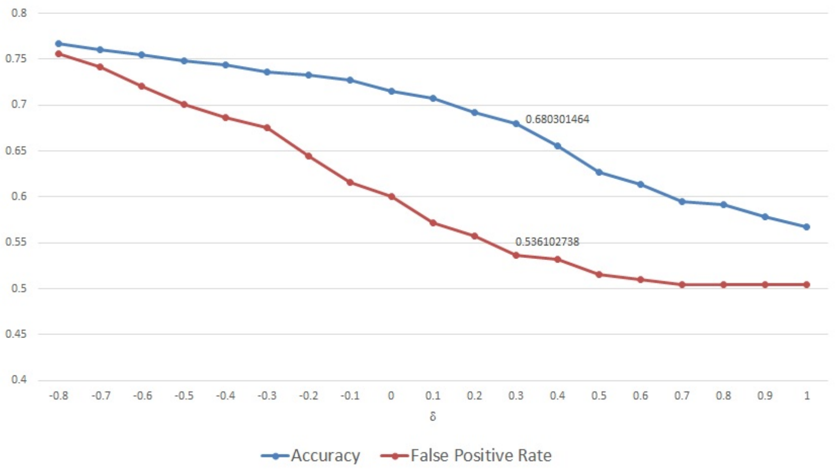

Figure 1: Characteristics of proposed markov model by varying the value of δ

Evaluation methods

To ensure the robustness of performance estimates, we used 10-fold cross validation for the annotation of clinical interview fragments, and 5-fold cross validation for the classification of sequence of behavior codes 36 as an experimental design. The performance of different classifiers and feature sets was evaluated in terms of precision, recall, F1 score (F1), kappa measure and accuracy using weighted macro-averaging over folds.

Results

Experimental evaluation of automatic annotation of clinical interview fragments and their sequences by using machine learning and probabilistic model included several dimensions:

- determining the performance of classifiers on the codebooks of different size;
- determining the effectiveness of the proposed contextual and semantic features.
- determining the performance of markov model in conjunction with determining PPC sequences that are most likely to translate into CT.

Since clinical researchers typically annotate caregiver and adolescent sessions separately, we first created two experimental datasets consisting of only adolescent and only caregiver session transcripts. Second, besides evaluating the accuracy of annotating adolescent and caregiver transcripts with the codebooks containing an entire set of codes, we also conducted a series of experiments with the codebooks of smaller sizes created as outlined above. Third, besides training and testing NB, SVM, CRF, Decision Tree, Boosting, DiscLDA, Random Forest and CNN classifiers using only lexical features, we also evaluated the effectiveness of the proposed contextual and semantic features.

Depending on the type of the interview transcript and the codebook size, SVM-AF achieves 3%–9% higher accuracy and 4%–10% higher F1 score than SVM and 4%–10% higher accuracy and 4%–11% higher F1 score than CRF, which highlights the importance of contextual and semantic features.

Performance of Markov model in conjunction with determining PPC code sequences that are most likely translate into CT

Since clinical researchers tried to increase the desire and ability of adolescent to change their current behavior to target behavior, we used sequence of successful and unsuccessful codes only from adolescent session for our sequential analysis. Summary of performance for the classification of sequential data is presented in Table **??**. From Table **??**, it follows that our designed markov model works very well in terms of precision and achieves precision 0.7574 with F1-measure 0.7092. The strength of the model for the classification of each type of sequences is illustrated by providing the confusion matrix in Table **??**. It also shows that successful patterns are identified more accurately compare to unsuccessful motivational interview sequences due to the imbalance of data.

Table 2: Performance of markov models for the classification of normal PPC code sequence.

Model	Order	Precision	Recall	F-Measure	Accuracy
General Markov Chain	First order	0.7387	0.7686	0.7532	0.7686
	Second order	0.6889	0.7889	0.7352	0.7889
Hidden Markov Model	First order	**0.7980**	0.8059	**0.7989**	0.8059
	Second order	0.7400	**0.8449**	0.7822	**0.8449**

Table 3: Performance of markov models for the classification of alternate PPC code sequence.

Model	Order	Precision	Recall	F-Measure	Accuracy
General Markov Chain	First order	0.9777	0.9437	0.9604	0.9437
	Second order	**0.9778**	0.9694	**0.9736**	0.9694
Hidden Markov Model	First order	0.9392	0.9313	0.9253	0.9392
	Second order	0.9704	**0.9713**	0.9699	**0.9713**

Two markov models are used to obtain top five most likely successful and unsuccessful motivational interviews that describing provider strategies and patient responses. Table **??** shows the top five successful and unsuccessful PPC code sequences.

Discussion

From the sequential analysis, we observed that markov model achieved near-human accuracy to categorise the sequence of behavior codes. It was also found that successful interviews are more frequently responded by the adolescent with PPC code 112. However, unsuccessful interviews are responded with the behavior code 109.

Conclusion

In this work, we suggest some successful sequences of codes for the practical use by the clinician during the interview session which will translate into positive change talk and commitment language. We also propose novel features and report the results of an extensive experimental evaluation of state-of-the-art supervised machine learning methods for text classification using those features, to help clinical researchers and practitioners assess the feasibility of using these methods for the task of automatic annotation of clinical text using the codebooks of realistic size. We found out that Support Vector Machine using only lexical features consistently outperforms all other classifiers on caregiver and adolescent datasets according to most metrics. Adding contextual and semantic features further improves the performance of SVM on both datasets, achieving close to human accuracy when the codebooks consisting of 16 and 17 classes are used to annotate caregiver and adolescent transcripts, respectively.

This work has important practical implications. First, it can facilitate researchers to establish causal relationship between different communication strategies and desired behavioral outcomes without having to repeatedly wade through pages of interview transcripts. Second, since automatic annotation is significantly faster than manual, it can dramatically accelerate the pace of research in behavioral sciences. Third, information that can directly inform and increase the efficiency of clinical practice for a successful interview. Although all experiments were conducted on interview transcripts, the proposed methods and features are not specific to a particular domain of Motivational Interviewing, and thus there is also no prima facie reason to believe that they will not be effective for annotation of any other type of clinical conversation.

Acknowledgments

We would like to thank the technical staff at Henry Ford Health Systems and Textual Data Analytics Laboratory at Wayne State University for their help with collecting and transforming binary data into free text EHR used for experiments reported in this paper.

References