Machine Learning Models for Text Segmentation

Mehedi Hasan, BS[1*], Alexander Kotov, PhD[1*], April Idalski Carcone, PhD[2], Ming Dong, PhD[1], Sylvie Naar, PhD[2]
[1]Department of Computer Science, Wayne State University, Detroit, Michigan
[2]Department of Family Medicine and Public Health Sciences, School of Medicine, Wayne State University, Detroit, Michigan

Abstract *The problem of analyzing temporally ordered observation sequences generated by a physiological, genomic or psychological process to make predictions about the outcome of that process arises in many domains of clinical informatics.*

Introduction

Temporally ordered sequences of discrete or continuous observations generated by genomic, psychological or psychological processes arise in many different domains of clinical informatics[1 2].

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 an MI interview session involving counselor, adolescent, and caregiver. The utterances were annotated based on MYSCOPE codebook[3], in which the codes are grouped into patient (adolescent and caregiver) codes and counselor codes. Utterances were divided into successful and unsuccessful communication sequences. Successful communication sequences result in positive change talk and commitment language statements by an adolescent or caregiver, 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. Successful sequences had an average length of 9.79 utterances, while unsuccessful sequences had on average 9.65 utterances. For each of the probabilistic models (MC and HMM), two models were trained, one model was trained using successful sequences and another one was trained using unsuccessful sequences.

Prediction method

Generally, a sequence can be viewed as a temporally ordered set of events. In this study, an event is a behavior code that also has a symbolic representation.

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 80% of the data and validated on 10%. The remaining 10% of the data is used as a test set for reporting the performance of the model.

Results

Experimental evaluation of the proposed method is conducted on both under and over-sampled sequences. Predictive performance summary of the proposed methods on under and over-sampled sequences is presented in Table.

[*] Authors provided equal contribution.

Discussion

By analyzing the experimental results of different communication sequence outcome prediction methods proposed in this paper, we arrived at the following conclusions. First, the overall predictive performance of RNN models is substantially better than probabilistic models. In particular, the RNN-based method achieves near-human accuracy for predicting the

Conclusion

In this paper, we compared the accuracy of Recurrent Neural Networks with Markov Chain and Hidden Markov Model for the task of predicting the success of motivational interviews. We found out that individual patient-provider communication exchanges

Acknowledgments

This study was supported by a grant from the National Institutes of Health, NIDDK R21DK108071, Carcone and Kotov, MPIs. 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 in developing the training dataset by manually annotating the dataset using the MYSCOPE codebook.

References

[1] 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.

[2] 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.

[3] 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.