Transformer-Based Observers in Psychotherapy

Tarun Sunkaraneni
University of Utah

UUCS-20-011

School of Computing
University of Utah
Salt Lake City, UT 84112 USA

7 August 2020

Abstract

Motivational Interviewing is a style of psychotherapy which has shown success as a method for treating addiction and substance abuse problems. In recent years, Natural Language Processing (NLP) techniques have shown promising results in assisting Motivational Interviewing (MI) training and advancing its popularity. One of the ways they are achieving this is by providing therapists feedback on their counseling efficacy by providing feedback to therapists via automatically tagged Motivational Interviewing skill codes (MISC) of utterances in a session.

The recent transformer architecture-based language models offer off-the-shelf, domain adaptation capabilities and have achieved state-of-the-art results in many NLP tasks. However, these models have not been trained on dialogue data, which is structured much differently than the linear structure of passages. We propose methods to encode dialogue data such that it can be domain adapted for psychotherapy. We achieve state-of-the art results in every measured metric of classifying Motivational Interviewing skill codes. While prior work has sought to model patient and therapist as separate agents, we show that a unified model can further improve state-of-the-art results. We conclude by examining how well classical NLP interpretation methods work on transformer models by analyzing probing results on these models. We are able to show that interpretation techniques, that have been useful for other NLP tasks such as question answering, may not be powerful enough to explain the inner workings of transformer models.