Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes

Jie Cao[†], **Michael Tanana**[‡], **Zac E. Imel**[‡], **Eric Poitras**[‡],
David C. Atkins[◇], **Vivek Srikumar**[†]
[†]School of Computing, University of Utah
[‡]Department of Educational Psychology, University of Utah
[◇]Department of Psychiatry and Public Health, University of Washington
{jcao, svivek}@cs.utah.edu,
{michael.tanana, zac.imel, eric.poitras}@utah.edu,
datkins@u.washington.edu

Abstract

Automatically analyzing dialogue can help understand and guide behavior in domains such as counseling, where interactions are largely mediated by conversation. In this paper, we study modeling behavioral codes used to asses a psychotherapy treatment style called Motivational Interviewing (MI), which is effective for addressing substance abuse and related problems. Specifically, we address the problem of providing real-time guidance to therapists with a dialogue observer that (1) categorizes therapist and client MI behavioral codes and, (2) forecasts codes for upcoming utterances to help guide the conversation and potentially alert the therapist. For both tasks, we define neural network models that build upon recent successes in dialogue modeling. Our experiments demonstrate that our models can outperform several baselines for both tasks. We also report the results of a careful analysis that reveals the impact of the various network design tradeoffs for modeling therapy dialogue.

1 Introduction

Conversational agents have long been studied in the context of psychotherapy, going back to chatbots such as ELIZA (Weizenbaum, 1966) and PARRY (Colby, 1975). Research in modeling such dialogue has largely sought to simulate a participant in the conversation.

In this paper, we argue for modeling dialogue *observers* instead of participants, and focus on psychotherapy. An observer could help an *ongoing* therapy session in several ways. First, by monitoring fidelity to therapy standards, a helper could guide both veteran and novice therapists towards better patient outcomes. Second, rather than generating therapist utterances, it could suggest the type of response that is appropriate. Third, it could alert a therapist about potentially important cues from a patient. Such assistance would be especially helpful in the increasingly prevalent online or text-based counseling services.[1]

We ground our study in a style of therapy called Motivational Interviewing (MI, Miller and Rollnick, 2003, 2012), which is widely used for treating addiction-related problems. To help train therapists, and also to monitor therapy quality, utterances in sessions are annotated using a set of behavioral codes called Motivational Interviewing Skill Codes (MISC, Miller et al., 2003). Table 1 shows standard therapist and patient (*i.e.*, client) codes with examples. Recent NLP work (Tanana et al., 2016; Xiao et al., 2016; Pérez-Rosas et al., 2017; Huang et al., 2018, *inter alia)* has studied the problem of using MISC to assess *completed* sessions. Despite its usefulness, automated post hoc MISC labeling does not address the desiderata for ongoing sessions identified above; such models use information from utterances yet to be said. To provide real-time feedback to therapists, we define two complementary dialogue observers:

1. **Categorization**: Monitoring an ongoing session by predicting MISC labels for therapist and client utterances as they are made.
2. **Forecasting**: Given a dialogue history, forecasting the MISC label for the next utterance, thereby both alerting or guiding therapists.

Via these tasks, we envision a helper that offers assistance to a therapist in the form of MISC labels.

We study modeling challenges associated with these tasks related to: (1) representing words and utterances in therapy dialogue, (2) ascertaining relevant aspects of utterances and the dialogue history, and (3) handling label imbalance (as evidenced in Table 1). We develop neural models that address these challenges in this domain.

Experiments show that our proposed models

[1]For example, Crisis Text Line (https://www.crisistextline.org), 7 Cups (https://www.7cups.com), etc.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5599–5611
Florence, Italy, July 28 - August 2, 2019.

Code	Count	Description	Examples
Client Behavioral Codes			
FN	47715	Follow/ Neutral: unrelated to changing or sustaining behavior.	"You know, I didn't smoke for a while." "I have smoked for forty years now."
CT	5099	Utterances about changing unhealthy behavior.	"I want to stop smoking."
ST	4378	Utterances about sustaining unhealthy behavior.	"I really don't think I smoke too much."
Therapist Behavioral Codes			
FA	17468	Facilitate conversation	"Mm Hmm.", "OK.","Tell me more."
GI	15271	Give information or feedback.	"I'm Steve.", "Yes, alcohol is a depressant."
RES	6246	Simple reflection about the clients most recent utterance.	C: "I didn't smoke last week" T: "Cool, you avoided smoking last week."
REC	4651	Complex reflection based on a client's history or the broader conversation.	C: "I didn't smoke last week." T: "You mean things begin to change".
QUC	5218	Closed question	"Did you smoke this week?"
QUO	4509	Open question	"Tell me more about your week."
MIA	3869	Other MI adherent,*e.g.*, affirmation, advising with permission, etc.	"You've accomplished a difficult task." "Is it OK if I suggested something?"
MIN	1019	MI non-adherent, *e.g.*, confrontation, advising without permission, etc.	"You hurt the baby's health for cigarettes?" "You ask them not to drink at your house."

Table 1: Distribution, description and examples of MISC labels.

outperform baselines by a large margin. For the categorization task, our models even outperform previous session-informed approaches that use information from future utterances. For the more difficult forecasting task, we show that even without having access to an utterance, the dialogue history provides information about its MISC label. We also report the results of an ablation study that shows the impact of the various design choices.[2].

In summary, in this paper, we (1) define the tasks of categorizing and forecasting Motivational Interviewing Skill Codes to provide real-time assistance to therapists, (2) propose neural models for both tasks that outperform several baselines, and (3) show the impact of various modeling choices via extensive analysis.

2 Background and Motivation

Motivational Interviewing (MI) is a style of psychotherapy that seeks to resolve a client's ambivalence towards their problems, thereby motivating behavior change. Several meta-analyses and empirical studies have shown the high efficacy and success of MI in psychotherapy (Burke et al., 2004; Martins and McNeil, 2009; Lundahl et al., 2010). However, MI skills take practice to master and require ongoing coaching and feedback to sustain (Schwalbe et al., 2014). Given the emphasis on using specific types of linguistic behaviors in MI (*e.g.*, open questions and reflections), fine-grained behavioral coding plays an important role in MI theory and training.

Motivational Interviewing Skill Codes (MISC, table 1) is a framework for coding MI sessions. It facilitates evaluating therapy sessions via utterance-level labels that are akin to dialogue acts (Stolcke et al., 2000; Jurafsky and Martin, 2019), and are designed to examine therapist and client behavior in a therapy session.[3]

As Table 1 shows, client labels mark utterances as discussing changing or sustaining problematic behavior (CT and ST, respectively) or being neutral (FN). Therapist utterances are grouped into eight labels, some of which (RES, REC) correlate with improved outcomes, while MI non-adherent (MIN) utterances are to be avoided. MISC labeling was originally done by trained annotators performing multiple passes over a session recording or a transcript. Recent NLP work speeds up this process by automatically annotating a completed MI session (*e.g.*, Tanana et al., 2016; Xiao et al., 2016; Pérez-Rosas et al., 2017).

Instead of providing feedback to a therapist after the completion of a session, can a dialogue observer provide online feedback? While past work has shown the helpfulness of post hoc eval-

[2]The code is available online at https://github.com/utahnlp/therapist-observer.

[3]The original MISC description of Miller et al. (2003) included 28 labels (9 client, 19 therapist). Due to data scarcity and label confusion, various strategies are proposed to merge the labels into a coarser set. We adopt the grouping proposed by Xiao et al. (2016); the appendix gives more details.

i	s_i	u_i	l_i
1	T:	Have you used drugs recently?	QUC
2	C:	I stopped for a year, but relapsed.	FN
3	T:	You will suffer if you keep using.	MIN
4	C:	Sorry, I just want to quit.	CT
...		...	...

Table 2: An example of ongoing therapy session

uations of a session, prompt feedback would be more helpful, especially for MI non-adherent responses. Such feedback opens up the possibility of the dialogue observer influencing the therapy session. It could serve as an assistant that offers suggestions to a therapist (novice or veteran) about how to respond to a client utterance. Moreover, it could help alert the therapist to potentially important cues from the client (specifically, CT or ST).

3 Task Definitions

In this section, we will formally define the two NLP tasks corresponding to the vision in §2 using the conversation in table 2 as a running example.

Suppose we have an ongoing MI session with utterances $u_1, u_2, \cdots, u_n$: together, the dialogue history H_n. Each utterance u_i is associated with its speaker s_i, either C (client) or T (therapist). Each utterance is also associated with the MISC label l_i, which is the object of study. We will refer to the last utterance u_n as the *anchor*.

We will define two classification tasks over a fixed dialogue history with n elements — *categorization* and *forecasting*. As the conversation progresses, the history will be updated with a sliding window. Since the therapist and client codes share no overlap, we will design separate models for the two speakers, giving us four settings in all.

Task 1: Categorization. The goal of this task is to provide real-time feedback to a therapist during an ongoing MI session. In the running example, the therapist's confrontational response in the third utterance is not MI adherent (MIN); an observer should flag it as such to bring the therapist back on track. The client's response, however, shows an inclination to change their behavior (CT). Alerting a therapist (especially a novice) can help guide the conversation in a direction that encourages it.

In essence, we have the following real-time classification task: *Given the dialogue history H_n which includes the speaker information, predict the MISC label l_n for the last utterance u_n.*

The key difference from previous work in predicting MISC labels is that we are restricting the input to the real-time setting. As a result, models can only use the dialogue history to predict the label, and in particular, we can not use models such as a conditional random field or a bi-directional LSTM that need both past and future inputs.

Task 2: Forecasting. A real-time therapy observer may be thought of as an expert therapist who guides a session with suggestions to the therapist. For example, after a client discloses their recent drug use relapse, a novice therapist may respond in a confrontational manner (which is not recommended, and hence coded MIN). On the other hand, a seasoned therapist may respond with a complex reflection (REC) such as *"Sounds like you really wanted to give up and you're unhappy about the relapse."* Such an expert may also anticipate important cues from the client.

The forecasting task seeks to mimic the intent of such a seasoned therapist: *Given a dialogue history H_n and the next speaker's identity s_{n+1}, predict the MISC code l_{n+1} of the yet unknown next utterance u_{n+1}.*

The MISC forecasting task is a previously unstudied problem. We argue that forecasting the type of the next utterance, rather than selecting or generating its text as has been the focus of several recent lines of work (*e.g.*, Schatzmann et al., 2005; Lowe et al., 2015; Yoshino et al., 2018), allows the human in the loop (the therapist) the freedom to creatively participate in the conversation within the parameters defined by the seasoned observer, and perhaps even rejecting suggestions. Such an observer could be especially helpful for training therapists (Imel et al., 2017). The forecasting task is also related to recent work on detecting antisocial comments in online conversations (Zhang et al., 2018) whose goal is to provide an early warning for such events.

4 Models for MISC Prediction

Modeling the two tasks defined in §3 requires addressing four questions: (1) How do we encode a dialogue and its utterances? (2) Can we discover discriminative words in each utterance? (3) Can we discover which of the previous utterances are relevant? (4) How do we handle label imbalance in our data? Many recent advances in neural networks can be seen as plug-and-play components. To facilitate the comparative study of models, we will describe components that address the above

questions. In the rest of the paper, we will use **boldfaced** terms to denote vectors and matrices and SMALL CAPS to denote component names.

4.1 Encoding Dialogue

Since both our tasks are classification tasks over a dialogue history, our goal is to convert the sequence of utterences into a single vector that serves as input to the final classifier.

We will use a hierarchical recurrent encoder (Li et al., 2015; Sordoni et al., 2015; Serban et al., 2016, and others) to encode dialogues, specifically a hierarchical gated recurrent unit (HGRU) with an utterance and a dialogue encoder. We use a bidirectional GRU over word embeddings to encode utterances. As is standard, we represent an utterance u_i by concatenating the final forward and reverse hidden states. We will refer to this utterance vector as $\boldsymbol{v}_i$. Also, we will use the hidden states of each word as inputs to the attention components in §4.2. We will refer to such contextual word encoding of the j^{th} word as $\boldsymbol{v}_{ij}$. The dialogue encoder is a unidirectional GRU that operates on a concatenation of utterance vectors $\boldsymbol{v}_i$ and a trainable vector representing the speaker s_i.[4] The final state of the GRU aggregates the entire dialogue history into a vector $\boldsymbol{H}_n$.

The HGRU skeleton can be optionally augmented with the word and dialogue attention described next. All the models we will study are two-layer MLPs over the vector $\boldsymbol{H}_n$ that use a ReLU hidden layer and a softmax layer for the outputs.

4.2 Word-level Attention

Certain words in the utterance history are important to categorize or forecast MISC labels. The identification of these words may depend on the utterances in the dialogue. For example, to identify that an utterance is a simple reflection (RES) we may need to discover that the therapist is mirroring a recent client utterance; the example in table 1 illustrates this. Word attention offers a natural mechanism for discovering such patterns.

We can unify a broad collection of attention mechanisms in NLP under a single high level architecture (Galassi et al., 2019). We seek to define attention over the word encodings $\boldsymbol{v}_{ij}$ in the history (called queries), guided by the word encodings in the anchor $\boldsymbol{v}_{nk}$ (called keys). The output is a sequence of attention-weighted vectors, one for each word in the i^{th} utterance. The j^{th} output vector $\boldsymbol{a}_j$ is computed as a weighted sum of the keys:

$$\boldsymbol{a}_{ij} = \sum_k \alpha_j^k \boldsymbol{v}_{nk} \tag{1}$$

The weighting factor α_j^k is the attention weight between the j^{th} query and the k^{th} key, computed as

$$\alpha_j^k = \frac{\exp\left(f_m(\boldsymbol{v}_{nk}, \boldsymbol{v}_{ij})\right)}{\sum_{j'} \exp\left(f_m(\boldsymbol{v}_{nk}, \boldsymbol{v}_{ij'})\right)} \tag{2}$$

Here, f_m is a match scoring function between the corresponding words, and different choices give us different attention mechanisms.

Finally, a combining function f_c combines the original word encoding $\boldsymbol{v}_{ij}$ and the above attention-weighted word vector $\boldsymbol{a}_{ij}$ into a new vector representation $\boldsymbol{z}_{ij}$ as the final representation of the query word encoding:

$$\boldsymbol{z}_{ij} = f_c(\boldsymbol{v}_{ij}, \boldsymbol{a}_{ij}) \tag{3}$$

The attention module, identified by the choice of the functions f_m and f_c, converts word encodings in each utterance $\boldsymbol{v}_{ij}$ into attended word encodings $\boldsymbol{z}_{ij}$. To use them in the HGRU skeleton, we will encode them a second time using a BiGRU to produce attention-enhanced utterance vectors. For brevity, we will refer to these vectors as $\boldsymbol{v}_i$ for the utterance u_i. If word attention is used, these attended vectors will be treated as word encodings.

To complete this discussion, we need to instantiate the two functions. We use two commonly used attention mechanisms: BiDAF (Seo et al.,

Method	f_m	f_c
BiDAF	$\boldsymbol{v}_{nk}\boldsymbol{v}_{ij}^T$	$[\boldsymbol{v}_{ij};\ \boldsymbol{a}_{ij};$ $\boldsymbol{v}_{ij} \odot \boldsymbol{a}_{ij};\ \boldsymbol{v}_{ij} \odot \boldsymbol{a}']$
GMGRU	$\boldsymbol{w}^e \tanh(\boldsymbol{W}^k \boldsymbol{v}_{nk} + \boldsymbol{W}^q[\boldsymbol{v}_{ij}; \boldsymbol{h}_{j-1}])$	$[\boldsymbol{v}_{ij}; \boldsymbol{a}_{ij}]$

Table 3: Summary of word attention mechanisms. We simplify BiDAF with multiplicative attention between word pairs for f_m, while GMGRU uses additive attention influenced by the GRU hidden state. The vector $\boldsymbol{w}_e \in \mathbb{R}^d$, and matrices $\boldsymbol{W}^k \in \mathbb{R}^{d \times d}$ and $\boldsymbol{W}^q \in \mathbb{R}^{2d \times 2d}$ are parameters of the BiGRU. The vector $\boldsymbol{h}_{j-1}$ is the hidden state from the BiGRU in GMGRU at previous position $j-1$. For combination function, BiDAF concatenates bidirectional attention information from both the key-aware query vector $\boldsymbol{a}_{ij}$ and a similarly defined query-aware key vector $\boldsymbol{a}'$. GMGRU uses simple concatenation for f_c.

[4] For the dialogue encoder, we use a unidirectional GRU because the dialogue is incomplete. For words, since the utterances are completed, we can use a BiGRU.

2016) and gated matchLSTM (Wang et al., 2017). For simplicity, we replace the sequence encoder in the latter with a BiGRU and refer to it as GMGRU. Table 3 shows the corresponding definitions of f_c and f_m. We refer the reader to the original papers for further details. In subsequent sections, we will refer to the two attended versions of the HGRU as BIDAF^H and GMGRU^H.

4.3 Utterance-level Attention

While we assume that the history of utterances is available for both our tasks, not every utterance is relevant to decide a MISC label. For categorization, the relevance of an utterance to the anchor may be important. For example, a complex reflection (REC) may depend on the relationship of the current therapist utterance to one or more of the previous client utterances. For forecasting, since we do not have an utterance to label, several previous utterances may be relevant. For example, in the conversation in Table 2, both u_2 and u_4 may be used to forecast a complex reflection.

To model such utterance-level attention, we will employ the multi-head, multi-hop attention mechanism used in Transformer networks (Vaswani et al., 2017). As before, due to space constraints, we refer the reader to the original work for details. We will use the $(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ notation from the original paper here. These matrices represent a query, key and value respectively. The multi-head attention is defined as:

$$\text{Multihead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = [\text{head}_1; \cdots; \text{head}_h]\boldsymbol{W}^O \quad (4)$$

$$\text{head}_i = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{W}_i^Q \left(\boldsymbol{K}\boldsymbol{W}_i^K\right)^T}{\sqrt{d_k}}\right)\boldsymbol{V}\boldsymbol{W}_i^V$$

The $\boldsymbol{W}_i$'s refer to projection matrices for the three inputs, and the final $\boldsymbol{W}^o$ projects the concatenated heads into a single vector.

The choices of the query, key and value defines the attention mechanism. In our work, we compare two variants: *anchor-based attention*, and *self-attention*. The anchor-based attention is defined by $Q = [\boldsymbol{v}_n]$ and $K = V = [\boldsymbol{v}_1 \cdots \boldsymbol{v}_n]$. Self-attention is defined by setting all three matrices to $[\boldsymbol{v}_1 \cdots \boldsymbol{v}_n]$. For both settings, we use four heads and stacking them for two hops, and refer to them as SELF$_{42}$ and ANCHOR$_{42}$.

4.4 Addressing Label Imbalance

From Table 1, we see that both client and therapist labels are imbalanced. Moreover, rarer labels are more important in both tasks. For example, it is important to identify CT and ST utterances. For therapists, it is crucial to flag MI non-adherent (MIN) utterances; seasoned therapists are trained to avoid them because they correlate negatively with patient improvements. If not explicitly addressed, the frequent but less useful labels can dominate predictions.

To address this, we extend the focal loss (FL Lin et al., 2017) to the multiclass case. For a label l with probability produced by a model p_t, the loss is defined as

$$\text{FL}(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t) \quad (5)$$

In addition to using a label-specific balance weight α_t, the loss also includes a modulating factor $(1-p_t)^\gamma$ to dynamically downweight well-classified examples with $p_t \gg 0.5$. Here, the α_t's and the γ are hyperparameters. We use FL as the default loss function for all our models.

5 Experiments

The original psychotherapy sessions were collected for both clinical trials and Motivational Interviewing dissemination studies including hospital settings (Roy-Byrne et al., 2014), outpatient clinics (Baer et al., 2009), college alcohol interventions (Tollison et al., 2008; Neighbors et al., 2012; Lee et al., 2013, 2014). All sessions were annotated with the Motivational Interviewing Skills Codes (MISC) (Atkins et al., 2014). We use the train/test split of Can et al. (2015); Tanana et al. (2016) to give 243 training MI sessions and 110 testing sessions. We used 24 training sessions for development. As mentioned in §2, all our experiments are based on the MISC codes grouped by Xiao et al. (2016).

5.1 Preprocessing and Model Setup

An MI session contains about 500 utterances on average. We use a sliding window of size $N = 8$ utterances with padding for the initial ones. We assume that we always know the identity of the speaker for all utterances. Based on this, we split the sliding windows into a client and therapist windows to train separate models. We tokenized and lower-cased utterances using spaCy (Honnibal and Montani, 2017). To embed words, we concatenated 300-dimensional Glove embeddings (Pennington et al., 2014) with ELMo vectors (Peters et al., 2018). The appendix details the model setup and hyperparameter choices.

5.2 Results

Best Models. Our goal is to discover the best client and therapist models for the two tasks. We identified the following best configurations using F_1 score on the development set:

1. **Categorization**: For client, the best model does not need any word or utterance attention. For the therapist, it uses GMGRU^H for word attention and ANCHOR$_{42}$ for utterance attention. We refer to these models as $\mathcal{C}_C$ and $\mathcal{C}_T$ respectively
2. **Forecasting**: For both client and therapist, the best model uses no word attention, and uses SELF$_{42}$ utterance attention. We refer to these models as $\mathcal{F}_C$ and $\mathcal{F}_T$ respectively.

Here, we show the performance of these models against various baselines. The appendix gives label-wise precision, recall and F_1 scores.

Results on Categorization. Tables 4 and 5 show the performance of the $\mathcal{C}_C$ and $\mathcal{C}_T$ models and the baselines. For both therapist and client categorization, we compare the best models against the same set of baselines. The majority baseline illustrates the severity of the label imbalance problem. Xiao et al. (2016), BiGRU$_{\text{generic}}$, Can et al. (2015) and Tanana et al. (2016) are the previous published baselines. The best results of previous published baselines are underlined. The last row Δ in each table lists the changes of our best model from them. BiGRU$_{\text{ELMo}}$, CONCAT^C, GMGRU^H and BiDAF^H are new baselines we define below.

Method	macro	FN	CT	ST
Majority	30.6	**91.7**	0.0	0.0
Xiao et al. (2016)	50.0	87.9	32.8	29.3
BiGRU$_{\text{generic}}$	50.2	87.0	35.2	28.4
BiGRU$_{\text{ELMo}}$	52.9	87.6	**39.2**	32.0
Can et al. (2015)	44.0	91.0	20.0	21.0
Tanana et al. (2016)	48.3	89.0	29.0	27.0
CONCAT^C	51.8	86.5	38.8	30.2
GMGRU^H	52.6	89.5	37.1	31.1
BiDAF^H	50.4	87.6	36.5	27.1
$\mathcal{C}_C$	**53.9**	89.6	39.1	**33.1**
$\Delta = \mathcal{C}_C -$ score	+3.5	-2.1	+3.9	+3.8

Table 4: Main results on categorizing client codes, in terms of macro F_1, and F_1 for each client code. Our model $\mathcal{C}_C$ uses final dialogue vector H_n and current utterance vector v_n as input of MLP for final prediction. We found that predicting using MLP(H_n) + MLP(v_n) performs better than just MLP(H_n).

The first set of baselines (above the line) do not encode dialogue history and use only the current utterance encoded with a BiGRU. The work of Xiao et al. (2016) falls in this category, and uses a 100-dimensional domain-specific embedding with weighted cross-entropy loss. Previously, it was the best model in this class. We also re-implemented this model to use either ELMo or Glove vectors with focal loss.[5]

The second set of baselines (below the line) are models that use dialogue context. Both Can et al. (2015) and Tanana et al. (2016) use well-studied linguistic features and then tagging the current utterance with both past and future utterance with CRF and MEMM, respectively. To study the usefulness of the hierarchical encoder, we implemented a model that uses a bidirectional GRU over a long sequence of flattened utterance. We refer to this as CONCAT^C. This model is representative of the work of Huang et al. (2018), but was reimplemented to take advantage of ELMo.

For categorizing client codes, BiGRU$_{\text{ELMo}}$ is a simple but robust baseline model. It outperforms the previous best no-context model by more than 2 points on macro F_1. Using the dialogue history, the more sophisticated model $\mathcal{C}_C$ further gets 1 point improvement. Especially important is its improvement on the infrequent, yet crucial labels CT and ST. It shows a drop in the F_1 on the FN label, which is essentially considered to be an unimportant, background class from the point of view of assessing patient progress. For therapist codes, as the highlighted numbers in Table 5 show, only incorporating GMGRU-based word-level attention, GMGRU^H has already outperformed many baselines, our proposed model $\mathcal{F}_T$ which uses both GMGRU-based word-level attention and anchor-based multi-head multihop sentence-level attention can further achieve the best overall performance. Also, note that our models outperform approaches that take advantage of future utterances.

For both client and therapist codes, concatenating dialogue history with CONCAT^C always performs worse than the hierarchical method and even the simpler BiGRU$_{\text{ELMo}}$.

Results on Forecasting. Since the forecasting task is new, there are no published baselines to compare against. Our baseline systems essentially differ in their representation of dialogue history. The model CONCAT^F uses the same architecture

[5] Other related work in no context exists (e.g., Pérez-Rosas et al., 2017; Gibson et al., 2017), but they either do not outperform (Xiao et al., 2016) or use different data.

Method	macro	FA	RES	REC	GI	QUC	QUO	MIA	MIN
Majority	5.87	47.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Xiao et al. (2016)	59.3	94.7	50.2	48.3	71.9	68.7	80.1	54.0	6.5
$\text{BiGRU}_{\text{generic}}$	60.2	94.5	50.5	49.3	72.0	70.7	80.1	54.0	10.8
$\text{BiGRU}_{\text{ELMo}}$	62.6	94.5	51.6	49.4	70.7	72.1	80.8	57.2	24.2
Can et al. (2015)	-	94.0	49.0	45.0	74.0	72.0	81.0	-	-
Tanana et al. (2016)	-	94.0	48.0	39.0	69.0	68.0	77.0	-	-
CONCAT^C	61.0	94.5	54.6	34.3	73.3	73.6	81.4	54.6	22.0
GMGRU^H	64.9	94.9	**56.0**	54.4	**75.5**	**75.7**	**83.0**	**58.2**	21.8
BiDAF^H	63.8	94.7	55.9	49.7	75.4	73.8	80.7	56.2	24.0
$\mathcal{C}_T$	**65.4**	**95.0**	55.7	**54.9**	74.2	74.8	82.6	56.6	**29.7**
$\Delta = \mathcal{C}_T - \underline{\text{score}}$	+5.2	+0.3	+3.9	+3.8	+0.2	+2.8	+1.6	+2.6	+18.9

Table 5: Main results on categorizing therapist codes, in terms of macro F_1, and F_1 for each therapist code. Models are the same as Table 4, but tuned for therapist codes. For the two grouped MISC set MIA and MIN, their results are not reported in the original work due to different setting.

Method	Dev		Test			
	CT	ST	macro	FN	CT	ST
CONCAT^F	20.4	30.2	43.6	84.4	23.0	**23.5**
HGRU	19.9	31.2	**44.4**	85.7	**24.9**	22.5
GMGRU^H	19.4	30.5	44.3	87.1	23.3	22.4
$\mathcal{F}_C$	**21.1**	**31.3**	44.3	85.2	24.7	22.7

(a) Main results on forecasting client codes, in terms of F_1 for ST, CT on dev set, and macro F_1, and F_1 for each client code on the test set.

Method	Recall	F_1								
	R@3	macro	FA	RES	REC	GI	QUC	QUO	MIA	MIN
CONCAT^F	72.5	23.5	63.5	0.6	0.0	53.7	27.0	15.0	18.2	9.0
HGRU	76.0	28.6	71.4	12.7	**24.9**	58.3	28.8	5.9	**17.4**	9.7
GMGRU^H	76.6	26.6	**72.6**	10.2	20.6	58.8	27.4	6.0	8.9	7.9
$\mathcal{F}_T$	**77.0**	**31.1**	71.9	**19.5**	24.7	**59.2**	**29.1**	**16.4**	15.2	**12.8**

(b) Main results on forecasting therapist codes, in terms of Recall@3, macro F_1, and F_1 for each label on test set

Table 6: Main results on forecasting task

as the model CONCAT^C from the categorizing task. We also show comparisons to the simple HGRU model and the GMGRU^H model that uses a gated matchGRU for word attention.[6]

Tables 6 (a,b) show our forecasting results for client and therapist respectively. For client codes, we also report the CT and ST performance on the development set because of their importance. For the therapist codes, we also report the recall@3 to show the performance of a suggestion system that displayed three labels instead of one. The results show that even without an utterance, the dialogue history conveys signal about the next MISC label. Indeed, the performance for some labels is even better than some categorization baseline systems. Surprisingly, word attention (GMGRU^H) in Table 6 did not help in forecasting setting, and a model with the SELF_{42} utterance attention is sufficient. For the therapist labels, if we always predicted the three most frequent labels (FA, GI, and RES), the recall@3 is only 67.7, suggesting that our models are informative if used in this suggestion-mode.

6 Analysis and Ablations

This section reports error analysis and an ablation study of our models on the development set. The appendix shows a comparison of pretrained domain-specific ELMo/glove with generic ones and the impact of the focal loss compared to simple or weighted cross-entropy.

6.1 Label Confusion and Error Breakdown

Figure 1 shows the confusion matrix for the client categorization task. The confusion between FN and CT/ST is largely caused by label imbalance. There are 414 CT examples that are predicted as ST and 391 examples vice versa. To further understand their confusion, we selected 100 of each for manual analysis. We found four broad categories of confusion, shown in Table 7.

[6]The forecasting task bears similarity to the next utterance selection task in dialogue state tracking work (Yoshino et al., 2018). In preliminary experiments, we found that the Dual-Encoder approach used for that task consistently underperformed the other baselines described here.

Category and Explaination	Client Examples (Gold MISC)
Reasoning is required to understand whether a client wants to change behavior, even with full context (50,42)	T: On a scale of zero to ten how confident are you that you can implement this change ? C: I don't know, seven maybe (CT); I have to wind down after work (ST)
Concise utterances which are easy for humans to understand, but missing information such as coreference, zero pronouns (22,31)	I mean I could try it (CT) Not a negative consequence for me (ST) I want to get every single second and minute out of it(CT)
Extremely short ($\leq$ 5) or long sentence ($\geq$ 40), caused by incorrect turn segementation. (21,23)	It is a good thing (ST) Painful (CT)
Ambivalent speech, very hard to understand even for human. (7,4)	What if it does n't work I mean what if I can't do it (ST) But I can stop whenever I want(ST)

Table 7: Categorization of CT/ST confusions.The two numbers in the brackets are the count of errors for predicting CT as ST and vice versa. We exampled 100 examples for each case.

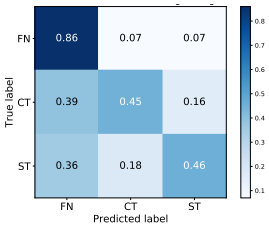

Figure 1: Confusion matrix for categorizing client codes, normalized by row.

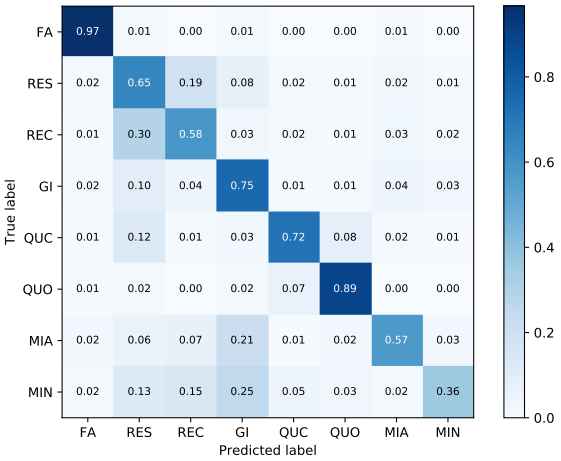

Figure 2: Confusion matrix for categorizing therapist codes, normalized by row.

The first category requires more complex reasoning than just surface form matching. For example, the phrase *seven out of ten* indicates that the client is very confident about changing behavior; the phrase *wind down after work* indicates, in this context, that the client drinks or smokes after work. We also found that the another frequent source of error is incomplete information. In a face-to-face therapy session, people may use concise and effient verbal communication, with guestures and other body language conveying information without explaining details about, for example, coreference. With only textual context, it is difficult to infer the missing information. The third category of errors is introduced when speech is transcribed into text. The last category is about ambivalent speech. Discovering the real attitude towards behavior change behind such utterances could be difficult, even for an expert therapist.

Figures 1 and 2 show the label confusion matrices for the best categorization models. We will examine confusions that are not caused purely by a label being frequent. We observe a common confusion between the two reflection labels, REC and RES. Compared to the confusion matrix from Xiao et al. (2016), we see that our models show much-decreased confusion here. There are two reason for this confusion persisting. First, the reflections may require a much longer information horizon. We found that by increasing the window size to 16, the overall reflection results improved. Second, we need to capture richer meaning beyond surface word overlap for RES. We found that complex reflections usually add meaning or emphasis to previous client statements using devices such as analogies, metaphors, or similes rather than simply restating them.

Closed questions (QUC) and simple reflections (RES) are known to be a confusing set of labels. For example, an utterance like *Sounds like you're suffering?* may be both. Giving information (GI) is easily confused with many labels because they relate to providing information to clients, but with different attitudes. The MI adherent (MIA) and non-adherent (MIN) labels may also provide information, but with supportive or critical attitude that may be difficult to disentangle, given the limited

Ablation	Options	macro	FN	CT	ST
history window size	0	51.6	87.6	39.2	32.0
	4	52.6	88.5	37.8	31.5
	8*	53.9	89.6	39.1	33.1
	16	52.0	89.6	39.1	33.1
word attention	+ GMGRU	52.6	89.5	37.1	31.1
	+ BiDAF	50.4	87.6	36.5	27.1
sentence attention	+ SELF$_{42}$	53.9	89.2	39.1	33.2
	+ ANCHOR$_{42}$	53.0	88.2	38.9	32.0

Table 8: Ablation study on categorizing client code. $*$ is our best model $\mathcal{C}_C$. All ablation is based on it. The symbol $+$ means adding a component to it. The default window size is 8 for our ablation models in the word attention and sentence attention parts.

Ablation	Options	macro	RES	REC	MIN
history window size	0	62.6	51.6	49.4	24.2
	4	64.4	54.3	53.2	23.7
	8*	65.4	55.7	54.9	29.7
	16	**65.6**	55.4	**56.7**	26.7
word attention	- GMGRU	62.0	51.9	51.7	16.0
	\ BiDAF	63.5	54.2	51.3	22.6
sentence attention	- ANCHOR$_{42}$	64.9	56.0	54.4	21.8
	\ SELF$_{42}$	63.4	55.5	48.2	21.1

Table 9: Ablation study on categorizing therapist codes, $*$ is our proposed model $\mathcal{C}_T$. \ means substituting and $-$ means removing that component. Here, we only report the important REC, RES labels for guiding, and the MIN label for warning a therapist.

number of examples.

6.2 How Context and Attention Help?

We evaluated various ablations of our best models to see how changing various design choices changes performance. We focused on the context window size and impact of different word level and sentence level attention mechanisms. Tables 8 and 9 summarize our results.

History Size. Increasing the history window size generally helps. The biggest improvements are for categorizing therapist codes (Table 9), especially for the RES and REC. However, increasing the window size beyond 8 does not help to categorize client codes (Table 8) or forecasting (in appendix).

Word-level Attention. Only the model $\mathcal{C}_T$ uses word-level attention. As shown in Table 9, when we remove the word-level attention from it, the overall performance drops by 3.4 points, while performances of RES and REC drop by 3.3 and 5 points respectively. Changing the attention to BiDAF decreases performance by about 2 points (still higher than the model without attention).

Sentence-level Attention. Removing sentence attention from the best models that have it decreases performance for the models $\mathcal{C}_T$ and $\mathcal{F}_T$ (in appendix). It makes little impact on the $\mathcal{F}_C$, however. Table 8 shows that neither attention helps categorizing clients codes.

6.3 Can We Suggest Empathetic Responses?

Our forecasting models are trained on regular MI sessions, according to the label distribution on Table 1, there are both MI adherent or non-adherent data. Hence, our models are trained to show how the therapist usually respond to a given statement.

To show whether our model can mimic *good* MI policies, we selected 35 MI sessions from our test set which were rated 5 or higher on a 7-point scale empathy or spirit. On these sessions, we still achieve a recall@3 of 76.9, suggesting that we can learn good MI policies by training on all therapy sessions. These results suggest that our models can help train new therapists who may be uncertain about how to respond to a client.

7 Conclusion

We addressed the question of providing real-time assistance to therapists and proposed the tasks of categorizing and forecasting MISC labels for an ongoing therapy session. By developing a modular family of neural networks for these tasks, we show that our models outperform several baselines by a large margin. Extensive analysis shows that our model can decrease the label confusion compared to previous work, especially for reflections and rare labels, but also highlights directions for future work.

Acknowledgments

The authors wish to thank the anonymous reviewers and members of the Utah NLP group for their valuable feedback. This research was supported by an NSF Cyberlearning grant (#1822877) and a GPU gift from NVIDIA Corporation.

References

David C Atkins, Mark Steyvers, Zac E Imel, and Padhraic Smyth. 2014. Scaling up the evaluation of psychotherapy: evaluating motivational interview-

ing fidelity via statistical text classification. *Implementation Science*, 9(1):49.

John S Baer, Elizabeth A Wells, David B Rosengren, Bryan Hartzler, Blair Beadnell, and Chris Dunn. 2009. Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors. *Journal of substance abuse treatment*, 37(2):191–202.

Brian L Burke, Christopher W Dunn, David C Atkins, and Jerry S Phelps. 2004. The emerging evidence base for motivational interviewing: A meta-analytic and qualitative inquiry. *Journal of Cognitive Psychotherapy*, 18(4):309–322.

Doğan Can, David C Atkins, and Shrikanth S Narayanan. 2015. A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Kenneth Mark Colby. 1975. *Artificial Paranoia: A Computer Simulation of Paranoid Process*. Pergamon Press.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2019. Attention, please! a critical review of neural attention models in natural language processing. *arXiv preprint arXiv:1902.02181*.

James Gibson, Dogan Can, Panayiotis Georgiou, David C Atkins, and Shrikanth S Narayanan. 2017. Attention networks for modeling behaviors in addiction counseling. In *Proceedings of the 2016 Conference of the International Speech Communication Association INTERSPEECH*.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Xiaolei Huang, Lixing Liu, Kate Carey, Joshua Woolley, Stefan Scherer, and Brian Borsari. 2018. Modeling temporality of human intentions by domain adaptation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 696–701.

Zac E Imel, Derek D Caperton, Michael Tanana, and David C Atkins. 2017. Technology-enhanced human interaction in psychotherapy. *Journal of counseling psychology*, 64(4):385.

Dan Jurafsky and James H Martin. 2019. *Speech and language processing*. Pearson.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

Christine M Lee, Jason R Kilmer, Clayton Neighbors, David C Atkins, Cheng Zheng, Denise D Walker, and Mary E Larimer. 2013. Indicated prevention for college student marijuana use: A randomized controlled trial. *Journal of consulting and clinical psychology*, 81(4):702.

Christine M Lee, Clayton Neighbors, Melissa A Lewis, Debra Kaysen, Angela Mittmann, Irene M Geisner, David C Atkins, Cheng Zheng, Lisa A Garberson, Jason R Kilmer, et al. 2014. Randomized controlled trial of a spring break intervention to reduce high-risk drinking. *Journal of consulting and clinical psychology*, 82(2):189.

Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. *arXiv preprint arXiv:1506.01057*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Ryan Lowe, Nissan Pow, Iulian V. Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of SIGDIAL*.

Brad W Lundahl, Chelsea Kunz, Cynthia Brownell, Derrik Tollefson, and Brian L Burke. 2010. A meta-analysis of motivational interviewing: Twenty-five years of empirical studies. *Research on social work practice*, 20(2):137–160.

Renata K Martins and Daniel W McNeil. 2009. Review of motivational interviewing in promoting health behaviors. *Clinical psychology review*, 29(4):283–293.

William Miller and Stephen Rollnick. 2003. Motivational interviewing: Preparing people for change. *Journal for Healthcare Quality*, 25(3):46.

William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. Manual for the motivational interviewing skill code (misc). *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*.

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

Clayton Neighbors, Christine M Lee, David C Atkins, Melissa A Lewis, Debra Kaysen, Angela Mittmann, Nicole Fossos, Irene M Geisner, Cheng Zheng, and Mary E Larimer. 2012. A randomized controlled trial of event-specific prevention strategies for reducing problematic drinking associated with 21st birthday celebrations. *Journal of consulting and clinical psychology*, 80(5):850.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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*, volume 1, pages 1128–1137.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Peter Roy-Byrne, Kristin Bumgardner, Antoinette Krupski, Chris Dunn, Richard Ries, Dennis Donovan, Imara I West, Charles Maynard, David C Atkins, Meredith C Graves, et al. 2014. Brief intervention for problem drug use in safety-net primary care settings: a randomized clinical trial. *Jama*, 312(5):492–501.

Jost Schatzmann, Kallirroi Georgila, and Steve Young. 2005. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *6th SIGdial Workshop on DISCOURSE and DIALOGUE*.

Craig S Schwalbe, Hans Y Oh, and Allen Zweben. 2014. Sustaining motivational interviewing: a meta-analysis of training studies. *Addiction (Abingdon, England)*, 109(8):1287–94.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. In *ICLR*.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, pages 3776–3784.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 553–562. ACM.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Michael Tanana, Kevin A Hallgren, Zac E Imel, David C Atkins, and Vivek Srikumar. 2016. A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of substance abuse treatment*, 65:43–50.

Sean J Tollison, Christine M Lee, Clayton Neighbors, Teryl A Neil, Nichole D Olson, and Mary E Larimer. 2008. Questions and reflections: the use of motivational interviewing microskills in a peer-led brief alcohol intervention for college students. *Behavior Therapy*, 39(2):183–194.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 189–198.

Joseph Weizenbaum. 1966. ELIZA – a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Bo Xiao, Dogan Can, James Gibson, Zac E Imel, David C Atkins, Panayiotis G Georgiou, and Shrikanth S Narayanan. 2016. Behavioral coding of therapist language in addiction counseling using recurrent neural networks. In *Proceedings of the 2016 Conference of the International Speech Communication Association INTERSPEECH*, pages 908–912.

Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, Chulaka Gunasekara, Walter S. Lasecki, Jonathan Kummerfeld, Michael Galley, Chris Brockett, Jianfeng Gao, Bill Dolan, Sean Gao, Tim K. Marks, Devi Parikh, and Dhruv Batra. 2018. The 7th dialog system technology challenge. *arXiv preprint*.

Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

A Appendix

Different Clustering Strategies for MISC The original MISC description of Miller et al. (2003) included 28 labels (9 client, 19 therapist). Due to data scarcity and label confusion, some labels were merged into a coarser set. Can et al. (2015) retain 6 original labels FA, GI, QUC, QUO, REC, RES, and merge remaining 13 rare labels into a

Code	Count	Description	Examples
MIA	3869	Group of MI Adherent codes : Affirm(AF); Reframe(RF); Emphasize Control(EC); Support(SU); Filler(FI); Advise with permission(ADP); Structure(ST); Raise concern with permission(RCP)	"You've accomplished a difficult task." (AF) "Its your decision whether you quit or not" (EC) "That must have been difficult." (SU) "Nice weather today!" (FI) "Is it OK if I suggested something?" (ADP) "Let's go to the next topic" (ST) "Frankly, it worries me." (RCP)
MIN	1019	Group of MI Non-adherent codes: Confront(CO); Direct(DI); Advise without permission(ADW); Warn(WA); Raise concern without permission(RCW)	"You hurt the baby's health for cigarettes?" (CO) "You need to xxx." (DI) "You ask them not to drink at your house." (ADW) "You will die if you don't stop smoking." (WA) "You may use it again with your friends." (RCW)

Table 10: Label distribution, description and exmaples for MIA and MIN

single COU label, they merge all 9 client codes into a single CLI label. Instead, Tanana et al. (2016) merge only 8 of rare labels into a OTHER label and they cluster client codes according to the valence of changing, sustaining or being neutral on the addictive behavior(Atkins et al., 2014). Then Xiao et al. (2016) combine and improve above two clustering strategies by splitting the all 13 rare labels according to whether the code represents MI-adherent(MIA) and MI-nonadherent (MIN) We show more details about the original labels in MIA and MIN in Table 10

Model Setup We use 300-dimensional Glove embeddings pre-trained on 840B tokens from Common Crawl (Pennington et al., 2014). We do not update the embedding during training. Tokens not covered by Glove are using a randomly initialized UNK embedding. We also use character-level deep contextualized embedding ELMo 5.5B model by concatenating the corresponding ELMo word encoding after the word embedding vector. For speaker information, we randomly initialize them with 8 dimensional vectors and update them during training. We used a dropout rate of 0.3 for the embedding layers.

We trained all models using Adam (Kingma and Ba, 2015) with learning rate chosen by cross validation between $[1e^{-4}, 5 * 1e^{-4}]$, gradient norms clipping from at $[1.0, 5.0]$, and minibatch sizes of 32 or 64. We use the same hidden size for both utterance encoder, dialogue encoder and other attention memory hidden size; it has been selected from $\{64, 128, 256, 512\}$. We set a smaller dropout 0.2 for the final two fully connected layers. All the models are trained for 100 epochs with early-stoping based on macro F_1 over development results.

Detailed Results of Our Main Models In the main text, we only show the F_1 score of each our proposed models. We summarize the performance of our best models for both categorzing and forecasting MISC codes in Table 11 with precision, recall and F_1 for each codes.

Label	Categorizing			Forecasting		
	P	R	F_1	P	R	F_1
FN	92.5	86.8	89.6	90.8	80.3	85.2
CT	34.8	44.7	39.1	18.9	28.6	22.7
ST	28.2	39.9	33.1	19.5	33.7	24.7
FA	95.1	94.7	94.9	70.7	73.2	71.9
RES	50.3	61.3	55.2	20.1	18.8	19.5
REC	52.8	55.5	54.1	19.2	34.7	24.7
GI	74.6	75.1	74.8	52.8	67.5	59.2
QUC	80.6	70.4	75.1	36.2	24.3	29.1
QUO	85.3	81.2	83.2	27.0	11.8	16.4
MIA	61.8	52.4	56.7	27.0	10.6	15.2
MIN	27.7	28.5	28.1	17.2	10.2	12.8

Table 11: Performance of our proposed models with respect to precision, recall and F_1 on categorizing and forecasting tasks for client and therapist codes

Domain Specific Glove and ELMo We use the general psychotherapy corpus with 6.5M words (Alexander Street Press) to train the domain specific word embeddings **Glove**$_{psyc}$ with 50, 100, 300 dimension. Also, we trained ELMo with 1 highway connection and 256-dimensional output size to get **ELMo**$_{psyc}$. We found that ELMo 5.5B performs better than ELMo psyc in our experiments, and general Glove-300 is better than the **Glove**$_{psyc}$. Hence for main results of our models, we use **ELMo**$_{generic}$ by default. Please see more details in Table 12

Model	Embedding	macro	FN	CT	ST	macro	FA	RES	REC	GI	QUC	QUO	MIA	MIN
$\mathcal{C}$	ELMo	53.9	89.6	**39.1**	**33.1**	**65.4**	**95.0**	**55.7**	**54.9**	**74.2**	**74.8**	**82.6**	**56.6**	**29.7**
	$\text{ELMo}_{\text{psyc}}$	46.9	88.9	27.5	24.3	64.2	94.9	53.3	53.3	75.8	74.8	82.2	56.1	23.5
	Glove	50.6	**89.9**	33.4	28.6	62.2	94.6	53.7	54.2	70.3	70.0	79.1	54.7	20.9
	$\text{Glove}^{\text{pysc}}$	47.4	88.4	23.9	30.0	63.4	94.9	54.7	52.8	75.2	71.4	80.8	53.6	23.5
$\mathcal{F}$	ELMo	**44.3**	**85.2**	**24.7**	22.7	**31.1**	71.9	19.5	**24.7**	**59.2**	28.3	**17.7**	15.9	9.0
	$\text{ELMo}_{\text{psyc}}$	43.8	84.0	22.4	25.0	29.1	**73.5**	15.5	24.3	59.1	**29.1**	9.5	12.1	10.1
	Glove	42.7	83.9	21.0	23.1	30.0	72.8	**20.8**	23.7	58.2	26.2	14.5	14.5	9.6
	$\text{Glove}^{\text{pysc}}$	43.6	81.9	23.3	**25.7**	30.8	72.1	19.7	24.4	57.3	28.9	13.7	**17.8**	**23.5**

Table 12: Ablation study for our proposed model with embeddings trained on the psychotherapy corpus.

Ablation	Options	CT	ST	R@3	FA	RES	REC	GI	QUC	QUO	MIA	MIN
history size	1	17.2	15.1	66.4	59.4	12.6	9.0	44.6	16.3	14.8	11.9	4.1
	4	16.8	22.6	75.3	71.4	15.6	21.1	57.1	**29.3**	11.0	11.2	14.4
	8^*	24.7	22.7	**77.0**	**72.8**	**20.8**	23.1	58.1	28.3	**17.7**	15.9	9.0
	16	23.9	20.7	76.5	71.2	13.7	24.1	**58.5**	25.9	9.7	16.2	12.7
word attention	GMGRU	14.0	**23.2**	75.7	71.7	14.2	23.0	57.5	26.5	8.0	15.4	11.6
	GMGRU_{4h}	19.1	22.9	76.3	71.3	12.1	23.3	58.1	24.5	12.6	11.7	14.0
sentence attention	$-\ \text{SELF}_{42}$	**24.9**	22.5	76.0	71.4	12.7	24.9	58.3	28.8	5.9	**17.4**	9.7
	$\setminus\ \text{ANCHOR}_{42}$	22.9	22.9	76.2	72.2	15.5	**24.6**	59.5	27.1	7.7	16.3	8.3
	$+\ \text{GMGRU} \setminus \text{ANCHOR}_{42}$	6.8	23.4	76.9	70.8	8.0	24.5	58.3	24.6	10.6	14.9	**12.1**

Table 13: Ablation on forecasting task on both client and therapist code. $*$ row are results of our best forecasting model $\mathcal{F}_C$, and $\mathcal{F}_T$. $\setminus$ means substitute anchor attention with self attention. +GMGRU ANCHOR_{42} means using word-level attention and achor-based sentence-level attention together.

Full Results for Ablation on Forecasting Tasks In addition to the ablation table in the main paper for categorizing tasks, we reported more ablation details on forecasting task in Table 13. Word-level attention shows no help for both client and therapist codes. While sentence-level attention helps more on therapist codes than on client codes. Multi-head self attention alsoachieves better performance than anchor-based attention in forecasting tasks.

Label Imbalance We always use the same α for all weighted focal loss. Besides considering the label frequency, we also consider the performance gap between previous reported F_1. We choose to balance weights α as {1.0,1.0,0.25} for CT,ST and FN respectively, and {0.5, 1.0, 1.0, 1.0, 0.75, 0.75,1.0,1.0} for FA, RES, REC, GI, QUC, QUO, MIA, MIN. As shown in Table 14, we report our ablation studies on cross-entropy loss, weighted cross-entropy loss, and focal loss. Besides the fixed weights, focal loss offers flexible hyperparameters to weight examples in different tasks. Experiments shows that except for the model $\mathcal{C}^T$, focal loss outperforms cross-entropy loss and weighted cross entropy.

Loss	Client			Therapist				
	F_1	CT	ST	F_1	RES	REC	MIA	MIN
$\mathcal{C}^{\text{ce}}$	47.0	28.4	22.0	60.9	54.3	53.8	53.7	4.8
$\mathcal{C}^{\text{wce}}$	53.5	39.2	32.0	65.4	55.7	54.9	56.6	29.7
$\mathcal{C}^{\text{fl}}$	53.9	39.1	33.1	65.4	55.7	54.9	56.6	29.7
$\mathcal{F}^{\text{ce}}$	42.1	17.7	18.5	26.8	3.3	20.8	16.3	8.3
$\mathcal{F}^{\text{wce}}$	43.1	20.6	23.3	30.7	17.9	25.0	17.7	10.9
$\mathcal{F}^{\text{fl}}$	44.2	24.7	22.7	31.1	19.5	24.7	15.2	12.8

Table 14: Abalation study of different loss function on categorizing and forecasting task. Based on our proposed model for our four settings, we compared our best model with crossentropy loss(ce), α balanced cross-entropy(wce) and focal loss. Here we only report the macro F_1 for rare labels and the overall macro F_1. $\gamma = 1$ is the best for both the model $\mathcal{C}_C$ and $\mathcal{F}_C$, while $\gamma = 0$ is the best for $\mathcal{C}_T$ and $\gamma = 3$ for $\mathcal{F}_T$. Worth to mention, when $\gamma = 0$, the focal loss degraded into α-balanced crossentropy, that first two rows are the same for therspit model.