Identifying therapist conversational actions across diverse psychotherapeutic approaches

Fei-Tzin Lee*, **Derrick Hull**†, **Jacob Levine**†, **Bonnie Ray**†, **Kathleen McKeown***
*Columbia University, Department of Computer Science
†Talkspace
{feitzin, kathy}@cs.columbia.edu
{derrick, bonnie.ray, jacob.levine}@talkspace.com

Abstract

While conversation in therapy sessions can vary widely in both topic and style, an understanding of the underlying techniques used by therapists can provide valuable insights into how therapists best help clients of different types. Dialogue act classification aims to identify the conversational "action" each speaker takes at each utterance, such as sympathizing, problem-solving or assumption checking. We propose to apply dialogue act classification to therapy transcripts, using a therapy-specific labeling scheme, in order to gain a high-level understanding of the flow of conversation in therapy sessions. We present a novel annotation scheme that spans multiple psychotherapeutic approaches, apply it to a large and diverse corpus of psychotherapy transcripts, and present and discuss classification results obtained using both SVM and neural network-based models. The results indicate that identifying the structure and flow of therapeutic actions is an obtainable goal, opening up the opportunity in the future to provide therapeutic recommendations tailored to specific client situations.

1 Introduction

Dialogue act classification is a task in which utterances in a conversation (or dialogue) are labeled with the *action* that utterance performs in the context of the dialogue - essentially, the intention of the speaker at that point in the conversation. In the general case, this might be something like a question, an agreement, or a backchannel, though the specific acts of interest depend on the application. This type of classification generally lends itself to a more thorough understanding of the flow of a conversation. For our application, psychotherapy, it can be particularly helpful in clarifying the specific patterns of behavior exhibited by the therapist in response to different client statements.

Mental health treatment is unique in that, unlike other specialties, intervention can take place directly through the interaction between a patient and the care provider or therapist (Gaut et al., 2017; Hull, 2014). This places critical emphasis on research to understand the dynamics and mechanisms of change within the interaction itself, just as medical investigators would perform for a newly advanced drug or surgical procedure. Historically, however, it has been too labor intensive to manually summarize sessions and therapist notes for record keeping, or to implement a process for reliably quantifying the flow and quality of the conversation, especially for large numbers of sessions or among large, heterogeneous samples. An automated avenue for labeling clinically relevant dialogue acts would allow us to learn patterns of discourse associated with differing clinical outcomes, potentially even uncovering patterns and effects that had previously remained hidden. The results could be used to inform the development of automated clinical assistants, conversational agents, and recommender or supervisory systems for therapists delivering care through technology.

In this paper we provide preliminary results towards this end on a dataset of therapy transcripts labeled with a novel set of high-level therapy-specific acts at the sentence level. While we are not at liberty to make the annotated corpus available publicly, we do include a description of the annotation scheme, and will release examples of our annotations. Our analyses result in two key findings: firstly, the *context* of the sentence provides the clearest and most stable signal of the act; and secondly, on our limited dataset, simple methods can achieve performance as good as or better than that of more complex approaches (i.e., our simple SVM classifier significantly outperformed more complex neural methods). We present a detailed error analysis of our models' performance on the development set

Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pages 12–23
Minneapolis, Minnesota, June 6, 2019.

to better understand where the approach works well and where it encounters the most challenges, and discuss future avenues of research to potentially address these challenges.

Our contributions include (1) a simple therapy-specific dialogue act classification scheme for therapist utterances relevant across a broad range of therapeutic approaches; (2) a sample of annotated utterances for a large corpus of diverse therapy transcripts; and (3) initial classification results on this dataset, with analysis.

2 Related Work

Several papers in recent years have developed machine learning approaches for the coding of dialogue in a psychotherapy context. Early work (Can et al., 2015) leveraged n-grams, dictionary-based features constructed based on psycho-linguistic norms such as LIWC (Pennebaker et al., 2015), and features used in more general dialog act classification modeling, such as that of (Jurafsky et al., 1997), to automate coding of therapist skill usage. More recent work has leveraged the methods of deep learning to incorporate the sequential aspects of client-therapist interactions, using variations on recurrent neural network models to improve the ability of the model to accurately classify therapist behaviors. See, for example, (Xiao et al., 2016; Gibson et al., 2016, 2017). This body of work has focused primarily on identifying therapist skills in Motivational Interviewing, a highly structured psychotherapy approach used for resolving ambivalence related to the treatment of conditions such as substance or alcohol use, or to engaging with treatment in general (Miller and Rollnick, 2012). Independently, Flemotomos et al. 2018 and Rojas Barahona et al. 2018 applied machine learning approaches to code behaviors common in the context of Cognitive Behavior Therapy (CBT). Rojas Barahona et al. developed neural network models for classification of various types of client 'thinking errors' identified as part of cognitive behavioral treatment, while Flemotomos et al. built SVM models to classify the overall quality of a CBT treatment session, looking at the distribution of different types of therapist behaviors used within the session, both process and content-oriented (e.g. homework assignments). CBT, while widely used, is again a fairly structured and goal-oriented approach to psychotherapy, making it more amenable to machine learning of underlying linguistic patterns. Other recent work (Gibson and Narayanan, 2018) has applied multi-task learning to transcripts representing both Motivational Interviewing and CBT-based approaches, an important advance due to the difficulty of obtaining large corpora of annotated transcripts for any single psychotherapy approach. Multi-label learning for concurrently classifying individual therapist utterances as well as the overall 'quality' of a session was also explored in the same paper.

Our work differs from these previous works in that our corpus of psychotherapy transcripts includes therapists using a variety of therapeutic approaches, including second- and third-wave CBT, psychodynamic, motivational interviewing, supportive/Rogerian, and an integrative or eclectic approach blending aspects of several approaches, thus providing less consistency in the language and behaviors exhibited by the therapists and making the automated coding task more difficult. To handle the greater heterogeneity of therapist speech, we have developed a broader annotation scheme that captures a wide variety of therapist behaviors common to the general therapeutic process, combining these with a small range of labels specific to particular approaches.

3 Data

3.1 Corpus

Our dataset consists of an annotated selection of transcripts from a corpus maintained by the publisher Alexander Street Press[1], available through library subscription; the full collection consists of approximately four thousand transcripts, 340 of which we labeled. In the base corpus, transcript lengths ranged from approximately 200 to 900 sentences. The client tended to speak more than the therapist, with client sentences ranging from 162 to 614 per transcript, while therapists spoke between 54 and 473 sentences (the entire dataset contained around 126,000 client sentences, and only 53,000 therapist sentences).

Transcripts were labeled with dialogue acts at the sentence level; some sentences were judged to contain no dialogue act in the annotation set and thus were left unlabeled. This left us with 8,420 labeled sentences from clients, and 9,056 labeled sentences from therapists. We focus on therapist act classification in this work, as it has proven easier

[1] https://alexanderstreet.com/products/counseling-and-psychotherapy-transcripts-series

Code	Description
Simple Reflection	Repeats client statement with minimal alteration.
Makes Needs Explicit	Identifies an implied or background need for the client.
Makes Emotions Explicit	Identifies an implied or background emotion for the client.
Makes Values Explicit	Identifies an implied or background value or set of values for the client.
Makes Relational Patterns Explicit	Identifies an implied or background relational pattern for the client.
Makes Consequences Explicit	Identifies an implied or background consequence of a client's action.
Makes Conflict Explicit	Identifies an implied or background emotional or situational conflict for the client.
Problem-Solving	Therapist offers possible solutions to a client problem.
Evokes Concrete Elaboration	More information about a specific event or statement is sought.
Evokes Perspective Elaboration	Client is asked to consider an experience from a different perspective or vantage point.
Narrowing	Therapist guides client to focus on a specific area of concern.
Planning	Therapist works with client to construct a specific plan of action.
Assumption Checking	Helps client determine if a thought or assumption is accurate or helpful.
Metaprocessing	Asks client to express how they are feeling in the immediate present about something that just happened in the therapy.
Makes Strengths/Resources Explicit	Identifies an implied or background strength or resource that the client exhibits.
Normalization	Client's experience is classified as "normal" or expectable by the therapist.
Sympathizing	Brief statements expressing regret for the challenges the client is having.
Reassuring	Therapist attempts to convince client that painful experiences are in fact okay or will get better.
Counterprojection	Makes assumptions the client might be making about the therapist or therapy explicit.
Teaching/Psychoeducation	Therapeutically relevant information about psychological principles is provided.
Self-Disclosure of Therapist Affect	Therapist expresses how they feel about what the client has said.

Table 1: Clinical codes for therapist. Sections indicate clinical codes in the categories Reflection, Question, Normalization/Misc, and Meta, in order.

both to define useful act categories and to practically classify acts for the therapist. Even though we are capturing several therapeutic approaches, therapists tend to deploy a limited range of dialogue acts and expressions, likely owing to the common elements among different psychotherapies and to shared aspects of clinical training and the clinical setting. Clients, on the other hand, are not operating from a handful of theoretical frameworks. They exhibit behavior that is less easy to organize and categorize, especially when drawing primarily on language.

3.2 Annotation scheme

To define the general section of the annotation scheme we drew from the dialogue acts identified in (Jurafsky et al., 1997) and selected those most pertinent to psychotherapy dialogue. The acts chosen were *Agreement, Disagreement, Apology, Thanking, Hedge, Opinion, Yes-No Question, Opening, Closing,* and *Signal Non-understanding*. These codes were used for both therapist and client. Clinical codes were identified for both therapist and client as relevant to psychotherapy and were derived from Emotion Focused Therapy (Pascual-Leone, 2018; Pascual-Leone and Greenberg, 2005), Cognitive Behavioral Therapy (Beck and Beck,

2011), Motivational Interviewing (Miller and Rollnick, 2012), and Accelerated Experiential Dynamic Psychotherapy (Fosha et al., 2009). There were 17 codes for client statements derived from the frameworks above and 21 therapist codes (see Table 1); when combined with the general codes, this resulted in 27 codes for the client and 31 for the therapist. As the client codes are not the focus of this work, we omit them from this paper. Therapist Statement codes are organized around whether the therapist is offering a statement to the client, making an observation, or emphasizing something in what the client said. Therapist Question codes cover the various kinds of questions or requests for more information that a therapist might invoke. Therapist codes were chosen that are determined by theory or previous research to be helpful, as well as those determined to be unhelpful. It is likely useful to identify both kinds of therapist behaviors for other clinical and analytic tasks.

3.3 Annotation process

A random sample of the total Alexander Street Corpus was annotated by 30 Masters level counseling and clinical psychology trainees using a spreadsheet annotation tool we adapted from Microsoft Excel functions. Annotators were trained by a clinical psychology researcher and could confer with others and the researcher when unsure about a particular annotation. The implementation allowed annotators to see each statement within the context of the overall therapy session and to annotate each statement with an individual general code and/or a clinical code when applicable. Each statement could receive both a general or clinical code, but only one of each. Codes were designed to minimize conceptual overlap at the sentence level.

3.4 Category selection

As the act classes were extremely unbalanced (see section 3.5) and due to annotator reliability concerns (see section 3.6), we merged our act codes into higher level categories (see Table 2) that would be more stable and easier to classify, while still clinically meaningful. We ended up with five classes: agreement (consisting of only the general code Agreement); reflection (consisting of the first section of Table 1); question (the second section of Table 1, and the general codes Yes-No Question and Signal Non-Understanding); Normalization/Misc (the third section, as well as Disagreement, Apology, Hedge, Opinion, and Opening from the general codes); and Meta (the final section, and the general code Closing).

Category	Sentences
Agreement	1277
Reflection	4016
Question	3164
Normalization/Misc	1715
Meta	790

Table 2: Class sizes for categories.

3.5 Data imbalance

Due to the already limited quantity of annotated data, we did not subsample classes to produce a balanced dataset. This resulted in a notable imbalance in our data, even at the category level, though much more so at the act level. Class sizes for categories are provided in table 2. Due to space constraints, we have left the class sizes at the act level for the appendix, but the largest act class for therapist was agreement, with 1277 samples, while there were nine classes with under a hundred samples.

3.6 Inter-annotator agreement

Agreement on low-level codes was fairly low for the client, though relatively high for the therapist: on the subset of sentences which were coded by two annotators, Cohen's kappa was 0.3164 for client sentences, and 0.7900 for therapist. Agreement on categories was higher: 0.6303 for client, and 0.8577 for therapist. Category agreement was computed by aggregating the total number of low-level acts that received a label within the category. The greater category-level agreement than act-level agreement indicates that most disagreements at the act level nevertheless fell within the same category - that is, for the same sentence, different annotators were more likely to mark two different act codes in the same category than they were to mark two different act codes corresponding to different categories altogether. Whether due to the complex and compound structure of certain sentences where multiple codes were possible or to the similar psychological function of different codes, the high-level categories appear to be more stable.

3.7 Data handling and preprocessing

Sentences were tokenized using the NLTK TweetTokenizer[2], with automatic lowercasing. In cases

[2]https://www.nltk.org/api/nltk.tokenize.html

where sentences had both a general and a clinical label, the clinical label was given precedence (i.e. the clinical label was used as the single "true" label). We used a 70/15/15% data split, yielding 6335, 1357, and 1359 sentences for our train, development and test sets, respectively.

4 Methods

4.1 Models

Our primary models include an SVM based on discrete features (n-grams, dialogue information, context features, and length) as well as two different neural network models - a feedforward neural net on the discrete features alone, and a convolutional neural network (CNN) over the text as well as the discrete features. For baselines, we used an SVM over n-grams only and a CNN over text only. In our initial experiments we also investigated recurrent models (RNNs), but found that convolutional models strongly outperformed these, and so we did not include an RNN in our final set of models.

4.2 Discrete features

We experimented with a number of different features, using n-grams from the sentence as our baseline. As features about the sentence itself, we included the length of the sentence (in tokens), as well as position information including the index of the sentence within the conversation (sentence position); as dialog features, we included the index of the speaker turn (turn position), and the index of the sentence within the current speaker turn (utterance position). As context-related information, we used labels from the immediate history of the sample sentence, with varying window sizes, as well as n-grams from those previous sentences. We also experimented with sentiment features for the sentence itself (minimum, maximum, and average word scores using SentiWordNet (Baccianella et al., 2010)); counts of words from two different psychologically meaningful dictionaries, LIWC (Pennebaker et al., 2015) and DAAP (Bucci and Maskit, 2005); part-of-speech tags; word embeddings; and metadata for the transcript. Of these, position and length information, context labels, and context n-grams provided a boost to performance over the baseline, and so we omitted the others from our final model. Thus, our final sets of features included sentence features (sentence position, length, and n-grams), context features (labels and n-grams), and dialogue features (speaker change, turn index, and sentence index within current turn). Interestingly, we found that using category-level labels as context labels provided better performance for category classification than using the more fine-grained act labels, perhaps due to therapists focusing on a particular approach, e.g. reflection, for multiple utterances in sequence before moving to a different type of intervention.

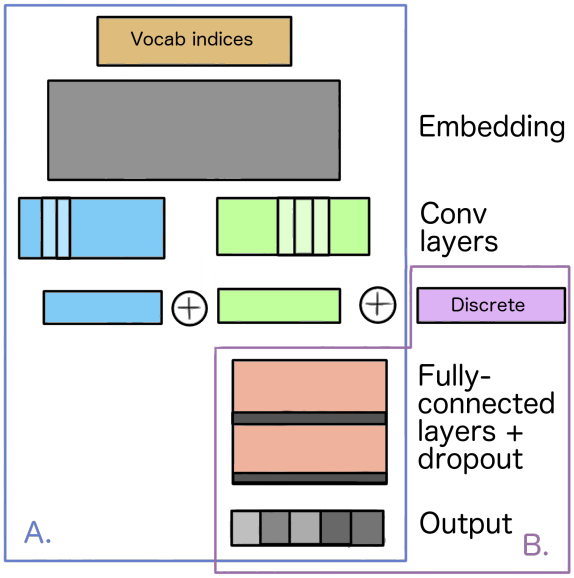

Figure 1: Architecture diagram for the full CNN model.

5 Experiments

5.1 Convolutional baseline

As our baseline model we use a convolutional neural network that takes as input only the text of the sentence and outputs a prediction in the form of a distribution over the category classes. We followed previous work (Liu et al., 2017) in the design of our architecture. The text is originally represented as a series of vocabulary indices; thus, the input to our model is initially a matrix whose dimensions are batch size (number of sentences) and sequence length (predefined number of words), where each element is a vocabulary index (see section A. of Figure 1). Sentences longer than the fixed maximum sequence length are clipped to that length, and shorter sentences are zero-padded. This array is passed through a 64-dimensional embedding layer with 0.5 dropout, followed by two parallel convolutional layers, one with window size 2 and one with window size 3. The representations produced by these two layers are concatenated and fed

into a series of two fully-connected dense layers with 0.5 dropout after each; our final layer performs softmax to produce the classification prediction. Intermediate layers use ReLU activation.

5.2 Other neural models

We experimented with two neural network architectures beyond the baseline. The first was a simple feedforward network running on the discrete features only (i.e. without word embeddings - see section B. of Figure 1), identical to the final component of the full architecture, consisting of two fully-connected layers with 128 nodes each, with dropout of 0.5 after each layer, and finally softmax over the classes.

The second was a convolutional net over the text combined with a feedforward component on the discrete features (see Figure 1). We used the same setup as the baseline, but concatenated the discrete features to the intermediate representations produced by the convolutional layers; the concatenated output was then processed by the fully-connected layers, mimicking the feedforward setup. Of our neural models, this latter model performed best.

5.3 Parameters and tuning

We performed gridsearch to find the optimal SVM parameters on different combinations of features. We found that a linear-kernel SVM performed best, with balanced class weights, l2 penalty, regularization parameter C = 0.01, and tolerance 0.3.

For the neural models, we used a batch size of 256, embedding dimension of 64, and maximum sequence length of 128 tokens; we trained for 16 epochs using Nadam optimization with .0002 learning rate, and crossentropy loss. We experimentally determined these parameters to be the best on the development set.

For the embedding layers in both convolutional nets we used random normal initialization and did not fix the weights, training the embedding weights along with the model parameters. Of the embedding initialization settings we tried (uniform random, random normal, and pretrained) this performed the best.

6 Results and Discussion

6.1 Category classification

Our evaluation task involved classifying individual sentences with one of the five act categories. Because of the high imbalance in class size, we used macro-F1 score as our primary statistic.

Classifier	Acc.	Pr.	Rc.	F1
Baseline SVM	70.20	61.77	60.75	60.27
Baseline CNN	49.99	28.26	36.39	29.60
Feedforward*†	74.07	**71.58**	65.96	67.66
CNN + features*†	74.52	70.61	66.68	68.00
SVM*†	**74.98**	70.91	**69.71**	**69.94**

Table 3: Classifier performance on test categories: accuracy, precision, recall and f-measure. Neural network results are reported as an average over five runs to account for variation in random initialization. (*) indicates significance over the SVM baseline, and (†) over the CNN baseline. More detailed results are presented in the appendix.

For all models, we experimented with feature selection, using Scikit-learn's SelectKBest feature selector, but found that reducing the number of features in this manner had a negative impact on development set performance. Thus, all final models equipped with discrete features used the full number of features. Although it seems likely that there would have been some uninformative features present in the large number (approximately 144,000) we ended up with, the lack of success of feature selection may be due to the small size of the training and validation sets, so that the features most informative on one may not have been the most informative on the other.

All final models performed significantly better than the baselines. Accuracy did not vary greatly between non-baseline classifiers (see Table 3). This is somewhat as expected - the majority classes were the easiest to classify, and classifiers performed well on them, while minority-class performance varied more but had less weight in the accuracy score. The other metrics (particularly recall and f-measure) showed more evident differences in performance, as they were weighted equally between classes. In overall performance, measured by macro-F1, the SVM was clearly the best. Interestingly, this was mostly due to a markedly higher recall than the neural methods, while its precision was between that of the feedforward net and the CNN. We used the Approximate Randomization Test (Riezler and Maxwell, 2005) to measure significance; oddly, the SVM achieved significance over every other method except the feedforward net. Considering that the SVM and feedforward net were the only two methods to receive exactly the same set of input features, this is perhaps due

	Text	SVM	True
1	Tell me your thoughts at that moment.	Meta	Question
2	So you've sort of ceased to mean all that much to him either?	Question	Reflection
3	Your mind really is just refusing to do it ... cause it doesn't want to and it's going to (inaudible).	Reflection	Reflection
4	Well, it's time for us to end but I guess I'm thinking ahead to the anniversary of your sister's death and I'm hoping that you get what you want.	Meta	Reflection

Table 4: Example classified sentences.

Field	Value	F1	
Therapy style	Client-centered therapy	71.29	1050
	Brief dynamic-relational therapy	48.96	201
	Experiential psychotherapy	58.78	65
	Cognitive behavioral therapy	84.17	41
Symptoms	Anger	65.86	430
	Anxiety	69.46	361
	Depression	71.13	322
	Low self-esteem	72.96	145
	Fearfulness	76.46	92
Therapist gender	Male	66.99	852
	Female	73.44	505

Table 5: Performance breakdown by metadata information on the development set. The final column contains the number of sentences present for the particular value of the specified field.

to some similarity in their outputs - possibly the feedforward net essentially performed as a slightly worse SVM, whereas the convolutional net had markedly different predictions, though with slightly better performance than the feedforward net.

6.2 Error analysis

In this section we analyze the performance of our best-performing model, the SVM with full feature sets. Agreement seemed easiest to classify, as one might expect; there were relatively few errors in that category. Unsurprisingly, the SVM tended to have difficulty with sentences that were requests for information not explicitly phrased as a question (e.g. example 1 in Table 4), as well as sentences phrased as questions that were not, in fact, questions - for instance, reflection-type rephrasings of the client's previous statement (example 2). Another major source of error was misclassification of normalization/misc statements as reflections. Both are similar in grammatical form and speak to the client's emotional experience. However, the intended psychological effect is different (reflections move to clarify and specify, normalizations act to reframe feelings in order to bring them down), and this difference was easy to miss or confuse. There was also a slight tendency to classify very short sentences as agreement, even if they were not - as agreement sentences are on average under four words per sentence, as opposed to most classes' 10-20, sentence length was a very strong signal for this class. On the other hand, the SVM was occasionally able to recover the labels of even sentences containing transcription artifacts such as (inaudible) or (ph) (see example 3).

One other quite interesting phenomenon we observed was that, upon close inspection, a number of the sentences that the SVM 'misclassified' in fact seemed to have been annotated incorrectly in the first place - for instance, example 4, which had been annotated as a reflection, but in fact should fall into the meta category, as the SVM predicted. This suggests the possibility of using a similar model as an annotation-checker of sorts, calling attention to sentences which coders might want to take a second or closer look at.

We also analyzed results across different therapy styles and other information about the transcript using the metadata available for the corpus (Table 5). One of the goals of the project was to develop a coding system capable of capturing important elements of several different therapies. The therapy style results suggest some progress in that direction. Interestingly, there was larger variation across therapy style than the other types of metadata. For

true/pred.	agr.	ref.	q.	misc	meta
agreement	153	7	1	3	2
reflection	20	444	62	22	25
question	4	50	302	9	14
norm/misc	3	53	3	55	9
meta	3	38	6	9	60

Table 6: Confusion matrix for SVM on development set categories.

example, accuracy for sentences taken from Brief Dynamic-Relational Therapy achieved an f-score of only 48.96 with the SVM, while Client-Centered Therapy had an f-score of 71.29. The SVM also did quite well with Cognitive Behavioral Therapy, but this class had only 41 samples. An examination of the annotated sentences for each therapy style themselves revealed two possible explanations for differences in accuracy. The first is that the sentences for Brief Dynamic-Relational and Experiential therapies tended to be nearly twice as long as those for Cognitive Behavioral therapies. They also tended to contain more comma splices and center embedding of clauses suggestive of more complex sentence structure. Secondly, the therapy styles with lower f-scores tended to have a smaller proportion of Agreement sentences (14% for Experiential and just 5% for Brief Dynamic-Relational compared to 46% for Cognitive Behavioral). The greater consistency in category distribution in these transcripts may have contributed to it being easier to guess the categories of their component sentences. Nevertheless, as there was generally very little data for each style, we presume that increasing the annotated data set for each style would help to diminish these differences and bring the therapy style f-scores closer together.

6.3 Ablation studies

From the final configuration of the SVM, we also performed ablation studies to determine which features had the most impact (Table 7). Context labels seemed to be by far the most important, with sentence n-grams second.

6.4 Negative results

In addition to the methods discussed here, we attempted a number of other techniques that were not successful (details presented in the appendix). To address the data scarcity issue, we pretrained on the Switchboard corpus; we tried a few different ways of distantly labeling the unlabeled data; we trained word embeddings on the unlabeled transcripts; we attempted to augment our dataset by "noising" sentences; and we attempted self-training with the unlabeled data. To address the discrepancy between reliability on act-level and category-level codes, we trained a cascading setup for the SVM, where a high-level classifier would first predict the category, and then the corresponding low-level classifier for that category would predict the act within that category. Finally, we attempted a basic weighted-average ensemble of our three non-baseline classifiers (SVM, feedforward net, and CNN with discrete features), as well as a more conservative ensemble that returned the SVM's prediction except when the SVM had low confidence, in which case it backed off to a weighted average.

Feature(s) removed	**p**	**r**	**f**
None	70	69.16	69.40
Sentence n-grams	64.64	66.80	65.50
Length	70.11	69.09	69.27
Sentence position	69.82	68.55	68.87
Context unigrams	69.67	68.43	68.43
Context labels	61.95	60.92	60.86
Speaker-change	69.85	68.95	69.2
Turn and intra-turn position	69.45	68.69	68.82

Table 7: Feature ablation for the SVM: precision, recall, and f-measure after removing features.

7 Conclusions and Future Work

We have created a new annotated corpus for therapy dialog act classification with labels at two levels of granularity, and analyzed classification results at each level. Our results indicate that context was very important, followed by sentence information, and that an SVM classifier is sufficient to make use of this information - our SVM model had significantly better performance than both the baselines and the neural methods we tried, aside from a feedforward net on exactly the same features.

One of the major challenges for this task was the limited size of the dataset. To address this, possible future directions include additional work on semisupervised learning, as well as an investigation into active learning for more efficient labeling. More broadly, future work might also focus more closely on the client's statements rather than only the therapist's, in order to glean a more comprehensive picture of the conversation.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA).

Judith S Beck and Aaron T Beck. 2011. *Cognitive Behavior Therapy*. Guilford Press, New York.

Wilma Bucci and Bernard Maskit. 2005. Building a weighted dictionary for referential activity. *Computing attitude and affect in text*, pages 49–60.

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*.

Nikolaos Flemotomos, Victor Martinez, James Gibson, David Atkins, Torrey Creed, and Shrikanth Narayanan. 2018. Language features for automated evaluation of cognitive behavior psychotherapy sessions. *Proc. Interspeech 2018*, pages 1908–1912.

Diana Fosha, Daniel J Siegel, and Marion Solomon. 2009. *The healing power of emotion: Affective neuroscience, development & clinical practice*. WW Norton & Company.

Garren Gaut, Mark Steyvers, Zac E Imel, David C Atkins, and Padhraic Smyth. 2017. Content coding of psychotherapy transcripts using labeled topic models. *IEEE journal of biomedical and health informatics*, 21(2):476–487.

James Gibson, Doğan Can, Panayiotis Georgiou, David C Atkins, and Shrikanth S Narayanan. 2017. Attention networks for modeling behaviors in addiction counseling. *Proc. Interspeech 2017*, pages 3251–3255.

James Gibson, Doğan Can, Bo Xiao, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth S Narayanan. 2016. A deep learning approach to modeling empathy in addiction counseling. *Interspeech 2016*, pages 1447–1451.

James Gibson and Shrikanth Narayanan. 2018. Multi-label multi-task deep learning for behavioral coding. *arXiv preprint arXiv:1810.12349*.

Thomas D Hull. 2014. Neuropsychiatric mhealth: Design strategies from emotion research. *mHealth Multidisciplinary Verticals*, page 199.

Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1997. Automatic detection of discourse structure for speech recognition and understanding. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 88–95. IEEE.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in dnn framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178. Association for Computational Linguistics.

William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford Press, New York.

A Pascual-Leone and LS Greenberg. 2005. Classification of affective-meaning states. *A. Pascual-Leone, Emotional processing in the therapeutic hour: Why the only way out is through*, pages 289–367.

Antonio Pascual-Leone. 2018. How clients change emotion with emotion: A programme of research on emotional processing. *Psychotherapy Research*, 28(2):165–182.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Stefan Riezler and John T Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for mt. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 57–64.

Lina M. Rojas Barahona, Bo-Hsiang Tseng, Yinpei Dai, Clare Mansfield, Osman Ramadan, Stefan Ultes, Michael Crawford, and Milica Gasic. 2018. Deep learning for language understanding of mental health concepts derived from cognitive behavioural therapy. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 44–54. Association for Computational Linguistics.

Anand Venkataraman, Andreas Stolcke, and Elizabeth Shriberg. 2002. Automatic dialog act labeling with minimal supervision.

Bo Xiao, Doğan Can, James Gibson, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth S Narayanan. 2016. Behavioral coding of therapist language in addiction counseling using recurrent neural networks. *Interspeech 2016*, pages 908–912.

Code	Samples	Wps
Agreement	1277	3.01
Disagreement	87	6.77
Apology	18	12.36
Thanking	7	7.4
Hedge	526	12.57
Opinion	676	14.19
Yes-no question	875	9.17
Signal non-understanding	215	8.52
Opening	63	8.30
Closing	90	6.74

Table 8: Distribution over general therapist act classes. "Wps" indicates the average number of words per sentence for that code.

A Code details

In this section we include more detailed statistics on the distribution of act-level classes in our data. Tables 8 and 9 include the number of sentences as well as the average number of words per sentence for each therapist act. The imbalance at the act level is far greater than that at the category level; the largest category is agreement, with 1277 sentences, while the smallest is thanking, with 7.

B Annotation process

A screenshot of the annotation spreadsheet is presented in Figure 2. Annotators were presented with a list of sentences and asked to choose an act or "u" (unlabeled) for each one.

Additionally, a confusion matrix for annotators' category labels is presented in Table 10. While the first annotator to give a label for each sentence was treated universally as "Annotator 1" and the second as "Annotator 2", not every sentence with two annotations was labeled by the same two annotators, and so this distinction is somewhat arbitrary. Nevertheless, this matrix still provides some notion of where disagreements occurred.

C Details of results

Further details of results are presented here. Table 11 contains performance broken down by category for the SVM classifier.

D Negative results

D.0.1 Distant labeling and data augmentation

As the most evident challenge with this dataset is the relatively small size - especially in the case of

Code	Samples	Wps
Simple reflection	638	9.10
Makes needs explicit	696	15.86
Makes emotions explicit	999	15.63
Makes values explicit	248	14.98
Makes relational patterns explicit	680	18.92
Makes consequences explicit	373	18.54
Makes conflict explicit	382	22.31
Makes strengths/resources explicit	122	18.01
Counterprojection	115	17.12
Teaching/psychoeducation	212	18.82
Problem-solving	166	16.93
Evokes concrete elaboration	1029	10.37
Evokes perspective flexibility	182	14.52
Narrowing	121	14.25
Planning	39	16.46
Assumption checking	426	14.77
Check in/metaprocessing	111	13.46
Self-disclosure	373	18.20
Normalization	77	17.15
Sympathizing	81	13.83
Reassuring	65	15.22

Table 9: Distribution over clinical therapist act classes. "Wps" indicates the average number of words per sentence for that code.

	agr.	refl.	q.	misc.	meta
agr.	46	0	0	5	0
refl.	0	136	24	7	6
q.	0	49	121	2	1
misc.	2	5	5	26	1
meta	1	3	0	0	11

Table 10: Annotator confusion matrix. Rows correspond to labels from the annotator 1, columns to labels from annotator 2.

category	precision	recall	F1
agreement	80.20	95.18	87.05
reflection	75.42	78.57	76.96
question	79.46	77.37	78.40
norm/misc	54.17	42.28	47.49
meta	65.31	55.17	59.81

Table 11: SVM performance by category.

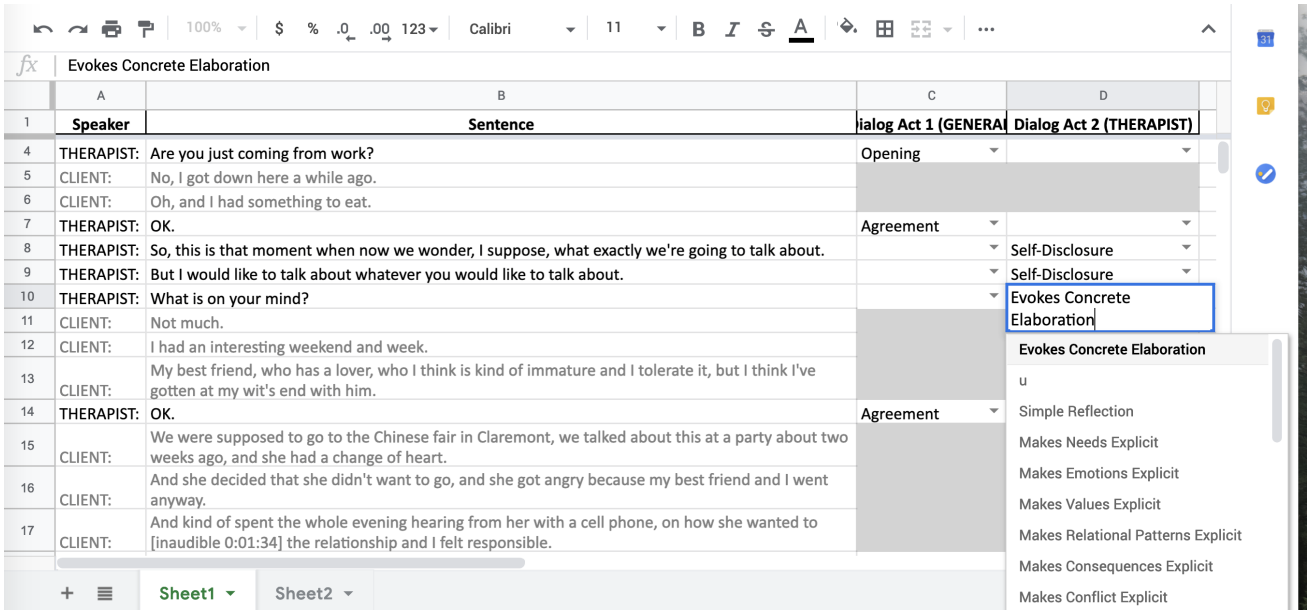

Figure 2: The interface that annotators used.

classification at the act level, in which the category classes are further subdivided - a natural course of inquiry was whether we could find additional data for transfer learning, produce noisy labels by some method on our much larger set of available unlabeled data, or leverage the unlabeled data in some other way.

Our first attempt in this direction was simply to add to our dataset the subset of labeled data from the Switchboard corpus corresponding to the labels that we had selected for our own annotation scheme. Surprisingly, this improved performance neither on the clinical labels nor even on the corresponding general labels. The fact that the Switchboard data was relatively uninformative for our own classification task suggests that the content of general-topic conversation (as in Switchboard) markedly differs from that found in therapy, as in our own corpus.

We next turned our attention to the remaining transcripts in the Alexander Street corpus that had not been labeled. We trained word embeddings on this data (using Word2Vec, with varying dimensionalities, and a window size of 7 and minimum count of 4); however, random initialization proved superior to both these and the publicly available pretrained embeddings trained on the Google News corpus.

As our SVM model had found success with relatively simple features, we also attempted to augment our dataset with distant labels generated by a few simple heuristic rules - if a sentence ends with '?', label it as a question; if it has relatively many agreement words, label it as agreement; return counterprojection if it has many "I" words (I, me, my, etc.); return normalization/misc if it has a high sentiment score; return reflection if it has many "you" words; and guess nothing otherwise.

Finally, observing the typical suite of tactics employed to boost the size and robustness of image datasets, we attempted to develop a similar technique for data augmentation in text. In essence, we drop or replace words randomly (with uniform probability, or with probability proportional to their smoothed unigram frequency). With a high base rate, this should produce highly noisy sentences that nevertheless contain some amount of signal approximating the original training data, hopefully improving classifier robustness. Unfortunately, this did not in fact improve performance.

D.0.2 Semisupervised learning

Partially inspired by the work of (Venkataraman et al., 2002), we explored self-training the SVM on sentences from the unlabeled transcripts. We experimented with a number of different learning schedules - adding all data labeled above a fixed confidence threshold to the training set in the next iteration; progressively increasing the confidence threshold by a fixed step at each iteration; halving the distance from the threshold to 100% confidence at each iteration; and scaling the base threshold by the ratio of current average confidence to original confidence over all unlabeled sentences at each iteration. Very small improvements were found under

some settings in preliminary work, but we did not explore this direction thoroughly as it yielded a dramatic increase in training time but only very minor gains in performance. Nevertheless, this might be worth revisiting in a more principled fashion in future work.

D.0.3 Ensembling

We attempted a couple simple methods of ensembling, in the hopes that our classifiers were different enough that this would yield useful information. The most basic of these was a simple weighted average of the prediction scores in each of the classes, with the highest averaged score being the final prediction. We also tried an ensemble-based method where we used the SVM's prediction unless its confidence was beneath a certain threshold, in which case we backed off to a weighted ensemble. Neither of these produced a performance improvement over the SVM; only the best weight assignment for classifiers that we found in the former case even approached the SVM's performance. This may be due to the high agreement between classifiers (agreement percentages between 86-92% for all three pairs of classifiers), meaning that none of them contributes new information relative to the others.

E Metadata analysis

We include breakdowns of performance by other metadata fields on the following page.

Psychological subject	F1	Samples
Emotional states	67.65	1458
Relationships	66.80	1244
Personality traits	70.10	516
Frustration	66.40	463
Spousal relationships	67.57	302
Behavior	74.20	285
Guilt	75.78	277
Family	64.35	267
Diagnosis	76.05	252
Sexual behavior	72.72	235
Communication	64.40	230
Client-counselor relations	64.36	193
Parent-child relationships	64.14	187
Personality factors	71.68	143
Ability	66.87	129
Self-confidence	67.19	97
Family relations	59.63	76

Table 12: Performance breakdown by psychological subject.

Experience	F1	Samples
Under 10 years	71.35	1102
11-20 years	75.35	118

Table 13: Performance breakdown by therapist experience.

Client age	F1	Samples
21-30 years	70.69	1200
31-40 years	77.13	40
41-50 years	42.94	109
51-60 years	54.17	8

Table 14: Performance breakdown by client age.

Client gender	F1	Samples
Male	69.22	744
Female	69.21	613

Table 15: Performance breakdown by client gender.