Leveraging Open Data and Task Augmentation to Automated Behavioral Coding of Psychotherapy Conversations in Low-Resource Scenarios

Zhuohao Chen[1], **Nikolaos Flemotomos**[1*], **Zac E. Imel**[2], **David C. Atkins**[3], **Shrikanth Narayanan**[1]
[1]University of Southern California, Los Angeles, CA, USA
[2]University of Utah, Salt Lake City, UT, USA
[3]University of Washington, Seattle, WA, USA
sail.usc.edu zac.imel@utah.edu datkins@u.washington.edu

Abstract

In psychotherapy interactions, the quality of a session is assessed by codifying the communicative behaviors of participants during the conversation through manual observation and annotation. Developing computational approaches for automated behavioral coding can reduce the burden on human coders and facilitate the objective evaluation of the intervention. In the real world, however, implementing such algorithms is associated with data sparsity challenges since privacy concerns lead to limited available in-domain data. In this paper, we leverage a publicly available conversation-based dataset and transfer knowledge to the low-resource behavioral coding task by performing an intermediate language model training via meta-learning. We introduce a task augmentation method to produce a large number of "analogy tasks" — tasks similar to the target one — and demonstrate that the proposed framework predicts target behaviors more accurately than all the other baseline models.

1 Introduction

Advances in spoken language processing techniques have improved the quality of life across several domains. One of the striking applications is automated behavioral coding in the fields of healthcare conversations such as psychotherapy. Behavioral coding is a procedure during which experts manually identify and annotate the participants' behaviors (Cooper et al., 2012). However, this process suffers from a high cost in terms of both time and human resources (Fairburn and Cooper, 2011). Building computational models for automated behavioral coding can significantly reduce the cost in time and provide scalable analytical insights into the interaction. A great amount of such work has been developed, including for addiction counseling (Tanana et al., 2016; Pérez-Rosas et al., 2017; Chen et al., 2019; Flemotomos et al., 2022) and couples therapy (Li et al., 2016; Tseng et al., 2016; Biggiogera et al., 2021). However, automated coding is associated with data sparsity due to the highly sensitive nature of the data and the costs of human annotation. Due to those reasons, *both* samples and labels of in-domain data are typically limited. This paper aims to train computational models for predicting behavior codes directly from psychotherapy utterances through classification tasks with limited in-domain data.

Recently, substantial work has shown the success of universal language representation via pre-training context-rich language models on large corpora (Peters et al., 2018; Howard and Ruder, 2018). Particularly, BERT (Bidirectional Encoder Representations from Transformers) has achieved state-of-the-art performance in many natural language processing (NLP) tasks and provided strong baselines in low-resource scenarios (Devlin et al., 2019). However, these models rely on self-supervised pre-training on a large out-of-domain text corpus. In prior works, the data sparsity issue has also been addressed by introducing an intermediate task pre-training using some other high-resource dataset (Houlsby et al., 2019; Liu et al., 2019; Vu et al., 2020). However, not all the source tasks yield positive gains. Sometimes the intermediate task might even lead to degradation due to the negative transfer (Pruksachatkun et al., 2020; Lange et al., 2021; Poth et al., 2021). To improve the chance of finding a good transfer source, we need to collect as many source tasks as possible. Another approach is meta-learning which aims to find optimal initialization for fine-tuning with limited target data (Gu et al., 2018; Dou et al., 2019; Qian and Yu, 2019). This approach also calls for enough source tasks and is affected by any potential task dissimilarity (Jose and Simeone, 2021; Zhou et al., 2021).

The challenge we need to handle is that both

*Work done while Nikolaos Flemotomos was at University of Southern California in 2022, and he is now affiliated to Apple Inc.

Code	Description	#Train	#Test
Therapist Utterances			
FA	Facilitate	19397	5838
GI	Giving information	17746	5064
RES	Simple reflection	7236	2137
REC	Complex reflection	4974	1510
QUC	Closed question	6421	1569
QUO	Open question	5011	1475
MIA	MI adherent	4898	1346
MIN	MI non-adherent	1358	237
Patient Utterances			
FN	Follow/Neutral	56204	15426
POS	Change talk	6146	1737
NEG	Sustain talk	5121	1407

Table 1: Data statistics for behavior codes in Motivational Interviewing psychotherapy.

utterances and assigned codes in psychotherapy interactions are domain-specific, making it difficult to leverage any open resource from a related domain. Considering that psychotherapy counseling takes place in a conversational setting, here we use a publicly available dataset — Switchboard-DAMSL (SwDA) corpus (Stolcke et al., 2000) — for the intermediate stages of knowledge transfer. Unlike most previous meta-learning frameworks, which require auxiliary tasks from various datasets, our work uses only one dataset and produces the source tasks by a task augmentation procedure. The task augmentation framework evaluates the correlations between the source and target labels. It produces source tasks by choosing subsets of source labels whose classes are in one-to-one correspondence with the target classes. Using this strategy, we can generate a large number of source tasks similar to the target task and thus improve the performance of meta-learning. The experimental results show that incorporating our proposed task augmentation strategy into meta-learning enhances the classification accuracy of automated behavioral coding tasks and outperforms all the other baseline approaches.

2 Dataset

We use data from Motivational Interviewing (MI) sessions of alcohol and drug abuse problems (Baer et al., 2009; Atkins et al., 2014) for the target task. The corpus consists of 345 transcribed sessions with behavioral codes annotated at the utterance level according to the Motivational Interviewing Skill Code (MISC) manual (Houck et al., 2010). We split the data into training and testing sets with a roughly 80%:20% ratio across speakers, resulting in 276 training sessions and 67 testing sessions. The statistics of the data are shown in Table 1.

We perform the intermediate task with the SwDA dataset, which consists of telephone conversations with a dialogue act tag for each utterance. We concatenate the parts of an interrupted utterance together, following Webb et al. (2005), which results in 196K training utterances and 4K testing utterances. This dataset supports 42 distinct tags, with more details displayed in Appendix A.

3 Methodology

3.1 Task Augmentation via Label Clustering

We define a low resource target task on $\mathcal{X} \times \mathcal{Y}$ and use $x \in \mathcal{X}$ to denote data and $y \in \mathcal{Y} = \{1, 2, ..., M\}$ to denote the target labels. We additionally assume a data-rich source task defined on $\mathcal{X} \times \mathcal{Z}$ with samples $\{(x_1, z_1), (x_1, z_2), ..., (x_n, z_n)\}$ supported by a much larger label set denoted by $z \in \mathcal{Z} = \{1, 2, ..., N\}$, $N > M$. Our task augmentation procedure aims at producing numerous tasks similar to the target task—we will refer to those as the *"analogy tasks"*.

The high-level idea is to construct the tasks with class labels similar to the target ones. Thus we explore the relationships between $\mathcal{Y}$ and $\mathcal{Z}$. We initialize M label subsets $C_1 = \varnothing, C_2 = \varnothing, \cdots, C_M = \varnothing$ to gather the source labels corresponding to $y = 1, y = 2, \cdots, y = M$, respectively. In the first step, we fine-tune on the in-domain target data to achieve a dummy classifier f. Then, we feed the source samples into f and obtain the predicted labels $\hat{Z} = \{f(x_1), f(x_2), \cdots, f(x_n)\}$. For any pair of a tar-

Algorithm 1 Construction of Analogy Tasks

Initialize model parameters θ; $K, M \in \mathbb{N}$.
Create empty label subsets: $C_1 = \varnothing, C_2 = \varnothing, ..., C_M = \varnothing$.
Fine-tune BERT with in-domain samples to obtain the classifier f
for i = 1 to K **do**
  for j = 1 to M **do**
    For the target label $y = j$, select $z^* \in \mathcal{Z}$ by Equation(1) and (2), then add it to C_j
    Remove the label z^* from $\mathcal{Z}$
Select one label from $C_1, C_2, ..., C_M$ to produce M^K analogy tasks

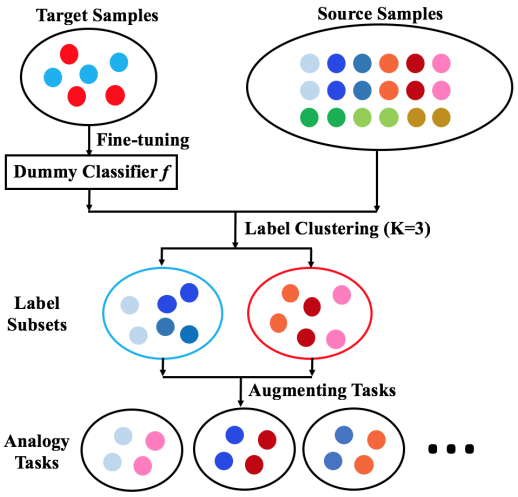

Figure 1: An example of task augmentation.

get label y and a source label z, we define the similarity function $Sim(\cdot)$ expressed by Equation (1). The value of $Sim(y, z)$ represents the proportion of the source samples within the class z, which are assigned the label y by f. For any target label y, we determine the most similar label z^* from the source data by Equation (2). Next, we apply Equations (1) and (2) to each of the target labels alternatively to cluster the source labels into the label subsets $C_1, C_2, \cdots, C_M$ with $|C_1| = |C_2| = \cdots = |C_M| = K$, where K is the size of the label subsets. Finally, we generate the source tasks by selecting one label from each of the subsets, resulting in M^K *analogy tasks*. The details of this procedure are given in Algorithm 1.

$$Sim(y, z) = \frac{\sum_{k=1}^{1} \mathbb{1}\{f(x_k) = y, z_k = z\}}{\sum_{i=1}^{1} \mathbb{1}\{z_k = z\}} \quad (1)$$

$$z^* = \arg\max_{z \in \mathcal{Z}} Sim(y, z) \quad (2)$$

Fig. 1 presents a task augmentation example from which we suppose the produced *analogy tasks* can benefit meta-learning in three aspects: 1) the task similarity and knowledge transfer are improved; 2) the large number of the *analogy tasks* increase the generalization which helps meta-learner find a commonly good model initialization; 3) the classification layers can be shared for all the tasks.

3.2 Meta-learning with Analogy Tasks

After task augmentation, we apply an optimization-based meta-learning algorithm for intermediate training with the produced analogy tasks $T_1, T_2, ..., T_{M^K}$. In particular, here we use Reptile, that has shown superior text classification results (Dou et al., 2019). We denote this Reptile-based framework with task augmentation as *Reptile-TA* and propose two task sampling methods:
Uniform, where we choose a task by uniformly selecting one source label from each label subset;
PPTS, where we choose an analogy task with the **p**robability **p**roportional to the **t**ask **s**ize to make the best use of instances (see Appendix B).

We describe the training procedure in Algorithm 2 where α and β present the learning rate for the inner and outer loop, respectively, and m denotes the update steps for the inner loop.

Algorithm 2 Reptile with Analogy Tasks

Initialize model parameters θ; $m \in \mathbb{N}$, $\alpha, \beta > 0$
for iteration in 1,2,... **do**
  Sample a batch of analogy tasks $\{\tau_i\}$ based on one of the sampling methods we proposed.
  for all τ_i **do**
    Compute θ_i^m by m gradient descent steps with the learning rate α.
    Update $\theta = \theta + \beta \frac{1}{|\{\tau_i\}|} \sum_{\tau_i} (\theta_i^m - \theta)$

4 Experimental Results

We adopt the MI dataset to perform two tasks: predicting the behavioral codes of the therapist and of the patient. We use SwDA as the source dataset for intermediate tasks to train the BERT model with Reptile. We set the number of sessions for both the training and validation sets to 1, 5, and 25 to simulate low-resource situations at different levels. We pick sessions randomly to make pairs of training and validation, and we repeat this 15 times. For each level of data sparsity, we report the averaged prediction results over 15 runs to reduce the effect of data variations.

4.1 Experimental Setup

Our BERT model was implemented in PyTorch (version 1.3.1) and initialized with **BERT-base**[1]. The model is trained using the Adam optimizer (Kingman and Ba, 2015) with a batch size of 64. In the Reptile stage, we set the learning rate to be α = 5e-5 for the inner loop and β = 1e-5 for the outer loop and fix the inner update step m to be 3 (Algorithm 2). We pre-train the model for 4 epochs and sample 8 tasks in each step. In the fine-tuning

[1]https://github.com/huggingface/pytorch-pretrained-BERT

Approach	Nb. Training Sessions		
	1	5	25
BERT	0.512	0.577	0.626
Pre-train-42	0.528	0.584	0.630
Pre-train-7	0.543	0.592	0.638
Pre-train-LC-Shared	0.533	0.584	0.633
Pre-train-LC-Unshared	0.552	0.597	0.643
Reptile-TA-Uniform	0.555	0.601	0.646
Reptile-TA-PPTS	**0.574**	**0.618**	**0.660**

Table 2: UARs on predicting therapist's codes.

Approach	Nb. Training Sessions		
	1	5	25
BERT	0.408	0.469	0.528
Pre-train-42	0.407	0.463	0.523
Pre-train-7	0.410	0.466	0.529
Pre-train-LC-Shared	0.445	0.497	0.545
Pre-train-LC-Unshared	0.446	0.499	0.545
Reptile-TA-Uniform	0.448	0.499	0.547
Reptile-TA-PPTS	**0.461**	**0.511**	**0.555**

Table 3: UARs on predicting patient's codes.

stage, we select the learning rate from {5e-6, 1e-5, 2e-5, 3e-5} and number of epochs from {1, 3, 5, 10} with the lowest validation loss via grid search. To handle the class imbalance, we assign a weight for each class inversely proportional to its class frequency in the fine-tuning stage. In the meta-learning stage, we assign a weight for each sample inversely proportional to the frequency of the label subsets it belongs. The tasks are evaluated by the unweighted average recall (UAR).

4.2 Baseline Methods

BERT: We directly fine-tune BERT with the limited in-domain data.

Pre-train-42: We pre-train the intermediate task of BERT with the SwDA dataset using a 42-class classification task adopting its standard label tags.

Pre-train-7: We cluster the labels into simpler 7 tags, as described by Shriberg et al. (1998), and pre-train the intermediate task of BERT with the SwDA dataset using a 7-class classification task.

To explore the effect of label clustering, we propose two more baseline approaches:

Pre-train-LC-Shared: After label clustering, we pre-train the model by classifying samples into the label subsets they belong to as in Fig. 1. The classification layer is shared between pre-training and fine-tuning stage.

Pre-train-LC-Unshared: The setup is the same as in **Pre-train-LC-Shared**, but the classification layer is randomly initialized for fine-tuning.

4.3 Results

The results of different algorithms for predicting therapist's and patient's codes are presented in Tables 2 and 3. For the therapist-related tasks, both *Pre-train-42* and *Pre-train-7* outperform fine-tuning BERT directly because some of the therapist's codes (i.e., "Open Question" or "Closed Question") are similar in function to dialog acts such as "Open Question" and "Yes-No Question". The *Pre-train-7* groups the source labels in a reasonable way, making the source task closer to the target task and achieving better performance than *Pre-train-42*. However, both failed to improve the accuracy of predicting patient behavior since the codes reflect whether the patient shows a motivation to change their behavior and thus do not have evident similarities to these dialogue acts. The results of *Pre-train-LC-Unshared* are better when compared to direct fine-tuning and regular pre-training. The greater improvement in the patient's task indicates that gathering the source labels similar to target labels is effective. However, sharing the classification layer when fine-tuning degrades the performance in the task of therapists. This drop is because the pre-trained models do not provide a good initialization, and thus, when we fine-tune BERT, it becomes difficult to escape from local minima.

The results under the dashed line in Tables 2 and 3 are for our proposed framework, where we set the size of label subsets K to be 3 and 8 for the therapist's task and patient's task, respectively. The outcomes show that our framework with task augmentation performs better than the baseline approaches. We further compare the performance using the different task sampling strategies proposed in Section 3.2, and the results demonstrate that *PPTS* is superior to *Uniform* achieving significantly better UAR scores than any other approaches at ($p < 0.05$) based on Student's t-test.

4.4 Effect of the Size of Label Subsets

We test the effect of K using *Pre-train-LC-Unshared* and *Reptile-TA-PPTS* with 5 training sessions. From the results in Tables 4 and 5 we find that an optimal K should be neither too small nor too big. When K is small, we utilize too little source data. A bigger value of K leads to a larger number of samples and augmented tasks. However,

Approach	Size of label subset K			
	2	3	4	5
Pre-train-LC-Unshared	0.585	**0.597**	0.596	0.589
Reptile-TA-PPTS	0.603	**0.618**	0.615	0.608

Table 4: Effect of the size of label subset K, 8-way classification tasks of therapist.

Approach	Size of label subset K			
	2	5	8	11
Pre-train-LC-Unshared	0.477	0.492	**0.499**	0.493
Reptile-TA-PPTS	0.488	0.501	**0.511**	0.504

Table 5: Effect of the size of label subset K, 3-way classification tasks of patient.

at the same time, it weakens the task similarity.

5 Conclusion and Future Work

This paper leveraged publicly available datasets to build computational models for predicting behavioral codes in psychotherapy conversations with limited samples. We employed a meta-learning framework with task augmentation based on the idea of analogy tasks to address the data limitation problem. We performed experiments at different sparsity levels and showed improvement over baseline methods. Besides, we discussed two task sampling strategies and the effect of a hyper-parameters in our framework. In the future, we plan to leverage contextual utterances into our algorithm and generalize our approach to the natural language understanding task in other fields. A more formal approach to find an optimal match between classes from different domains (i.e., labels of conversational descriptors) is also a topic of our ongoing research.

6 Limitations

As an initial stage in an ongoing effort, our work has several limitations. First, we only leverage a single open dataset for the intermediate task. There are other conversation-based corpora with utterance-level labels that we have not explored yet, such as Persuasion For Good Corpus (Wang et al., 2019) and DailyDialog Corpus (Li et al., 2017). Second, we adopted the **BERT-base** as the language model throughout all the experiments ignoring domain adaptation. For example, we can perform domain-adaptive pre-training with a publicly available general psychotherapy corpus (Imel et al., 2015). In our framework, we force the size of the label subsets to be the same in the label clustering stage, which might be sub-optimal. A more sophisticated clustering algorithm is needed. Besides, the intermediate task can introduce biases into the target, which calls for more discussion.

7 Ethical Considerations

As a research focused on psychotherapy and automated behavioral coding using speech and language processing techniques, it is necessary to review the ethical implications of this work.

Given the sensitive nature of the data, the primary ethical issue is the privacy of all the participating individuals - both patients and therapists. Informed consent was employed to make sure the recording is permitted by the participants, in adherence to professional guidelines (Association, 2002). All the researchers involved in the study are trained and certifies on human subject data research, and all the data are stored in dedicated secure machines with restricted access. It was guaranteed that these data will not be shared with anyone who is not involved in the study. The current study is governed by restrictions imposed by the relevant Institutional Review Board (IRB).

References

American Psychological Association. 2002. Ethical principles of psychologists and code of conduct. *American psychologist*, 57(12):1060–1073.

David C Atkins, Mark Steyvers, Zac E Imel, and Padhraic Smyth. 2014. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, 9(1):1–11.

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.

Jacopo Biggiogera, George Boateng, Peter Hilpert, Matthew Vowels, Guy Bodenmann, Mona Neysari, Fridtjof Nussbeck, and Tobias Kowatsch. 2021. Bert meets liwc: Exploring state-of-the-art language models for predicting communication behavior in couples' conflict interactions. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pages 385–389.

Zhuohao Chen, Karan Singla, James Gibson, Dogan Can, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth Narayanan. 2019. Improving the

prediction of therapist behaviors in addiction counseling by exploiting class confusions. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6605–6609. IEEE.

Harris Ed Cooper, Paul M Camic, Debra L Long, AT Panter, David Ed Rindskopf, and Kenneth J Sher. 2012. *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* American Psychological Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1192–1197, Hong Kong, China. Association for Computational Linguistics.

Christopher G Fairburn and Zafra Cooper. 2011. Therapist competence, therapy quality, and therapist training. *Behaviour research and therapy*, 49(6-7):373–378.

Nikolaos Flemotomos, Victor R Martinez, Zhuohao Chen, Karan Singla, Victor Ardulov, Raghuveer Peri, Derek D Caperton, James Gibson, Michael J Tanana, Panayiotis Georgiou, et al. 2022. Automated evaluation of psychotherapy skills using speech and language technologies. *Behavior Research Methods*, 54(2):690–711.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

JM Houck, TB Moyers, WR Miller, LH Glynn, and KA Hallgren. 2010. Motivational interviewing skill code (misc) version 2.5. *(Available from http://casaa.unm.edu/download/misc25.pdf)*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Zac E Imel, Mark Steyvers, and David C Atkins. 2015. Computational psychotherapy research: Scaling up the evaluation of patient–provider interactions. *Psychotherapy*, 52(1):19.

Sharu Theresa Jose and Osvaldo Simeone. 2021. An information-theoretic analysis of the impact of task similarity on meta-learning. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1534–1539. IEEE.

DP Kingman and J Ba. 2015. Adam: A method for stochastic optimization. conference paper. In *Proceedings of the 3rd International Conference on Learning Representations*.

Lukas Lange, Jannik Strötgen, Heike Adel, and Dietrich Klakow. 2021. To share or not to share: Predicting sets of sources for model transfer learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8744–8753, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haoqi Li, Brian Baucom, and Panayiotis Georgiou. 2016. Sparsely connected and disjointly trained deep neural networks for low resource behavioral annotation: Acoustic classification in couples' therapy. *Interspeech 2016*, pages 1407–1411.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence An, 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*, 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 *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. 2021. What to pre-train on? efficient intermediate task selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10585–10605.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Kun Qian and Zhou Yu. 2019. Domain adaptive dialog generation via meta learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2639–2649, Florence, Italy. Association for Computational Linguistics.

Elizabeth Shriberg, Andreas Stolcke, Daniel Jurafsky, Noah Coccaro, Marie Meteer, Rebecca Bates, Paul Taylor, Klaus Ries, Rachel Martin, and Carol Van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and speech*, 41(3-4):443–492.

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.

Shao-Yen Tseng, Sandeep Nallan Chakravarthula, Brian R Baucom, and Panayiotis G Georgiou. 2016. Couples behavior modeling and annotation using low-resource lstm language models. In *Interspeech*, pages 898–902.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926, Online. Association for Computational Linguistics.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Nick Webb, Mark Hepple, and Yorick Wilks. 2005. Dialogue act classification based on intra-utterance features. In *Proceedings of the AAAI Workshop on Spoken Language Understanding*, volume 4, page 5. Citeseer.

Pan Zhou, Yingtian Zou, Xiao-Tong Yuan, Jiashi Feng, Caiming Xiong, and Steven Hoi. 2021. Task similarity aware meta learning: Theory-inspired improvement on maml. In *Uncertainty in Artificial Intelligence*, pages 23–33. PMLR.

Appendix

A The SwDA Dataset

This section demonstrates the dialogue acts distributions of the SwDA dataset. The statistics for the 42-tag scheme and the simpler 7-tag scheme are presented in Tables 6 and 7, respectively.

B A proposition for the Sampling Strategy PPTS

Proposition 1 *If we adopt the sampling strategy PPTS, then every unique instance within the label subsets has the same chance of being picked.*

Proof. Define the label subsets after label clustering $C_1 = \{c_1^1, c_1^2, ..., c_1^K\}$, $C_2 = \{c_2^1, c_2^2, ..., c_2^K\}$,..., $C_M = \{c_M^1, c_M^2, ..., c_M^K\}$.

Let x be an arbitrary instance with label which is contained in the label subset C_i, $1 \leq i \leq M$.

Consider the following process: 1) sample an analogy task with the probability proportional to the task size; 2) randomly sample an instance from the selected task.

We compute the probability of the picked instance to be x by

$$\begin{aligned} P(x) &= \sum_T P(T) \cdot P(x|T) \\ &= \sum_T \frac{|T|}{\sum_T |T|} \cdot \frac{1}{|T|} \cdot \mathbb{1}\{x \in T\} \\ &= \sum_T \frac{\mathbb{1}\{x \in T\}}{\sum_T |T|} \end{aligned} \tag{3}$$

where T denotes any analogy task. Consider that an arbitrary label c can be enrolled in exact K^{M-1} analogy tasks. Equation 3 can be rewritten as

$$\begin{aligned} P(x) &= \sum_T \frac{K^{M-1}}{K^{M-1} \sum_i^M \sum_j^K |c_i^j|} \\ &= \frac{1}{\sum_i^M \sum_j^K |c_i^j|} \end{aligned} \tag{4}$$

The probability is irrelevant to the label and thus the same for every instance x. Please note that the proposition will not hold if the sizes of the label subsets C_i are different.

C Examples of Label Subsets

This part shows examples of label clustering results for predicting therapist's codes and patient's codes with five in-domain training sessions. Table 8 and 9 present the produced label subsets which achieve the median performance among 15 runs of *Reptile-TA-PPTS*.

Dialogue Act	Utterances (count)	Dialogue Act	Utterances (count)
statement-non-opinion	74k	collaborative completion	0.7k
backchannel	38k	repeat-phrase	0.7k
statement-opinion	26k	open question	0.6k
abandoned/uninterpretable	15k	rhetorical questions	0.6k
agree/accept	11k	hold-before-answer/agreement	0.5k
appreciation	4.7k	reject	0.3k
yes-no-question	4.7k	negative non-no answers	0.3k
non-verbal	3.6k	signal-non-understanding	0.3k
yes answers	3k	other answer	0.3k
Conventional-closing	2.6k	conventional-opening	0.2k
wh-question	1.9k	or-clause	0.2k
no answers	1.4k	dispreferred answers	0.2k
response acknowledgement	1.3k	3rd-party-talk	0.1k
hedge	1.2k	offers, options commits	0.1k
declarative yes-no-question	1.2k	self-talk	0.1k
backchannel in question form	1k	downplayer	0.1k
quotation	0.9k	maybe/accept-part	0.1k
summarize/reformulate	0.9k	tag-question	0.1k
other	0.9k	declarative wh-question	0.1k
affirmative non-yes answers	0.8k	apology	0.1k
action-directive	0.7k	thinking	0.1k

Table 6: Statistics describing the SwDA datasets for the 42 tags scheme.

Dialogue Act	Utterances (count)
statement	100k
backchannel	38k
question	8.6k
agreement	11k
appreciation	4.7k
incomplete	15k
other	23k

Table 7: Statistics describing the SwDA datasets for the 7 tags scheme (Shriberg et al., 1998).

Behavioral Code	Clustered Similar Labels
Facilitate	backchannel, yes answer, no answer
Giving Information	statement-opinion, statement-non-opinion, dispreferred-answers
Simple Reflection	quotation, declarative yes-no-question, declarative wh-question
Complex Reflection	non-verbal, hedge, summarize/reformulate
Closed Question	yes-no-question, or-clause, tag-question
Open Question	wh-question, open-question, self-talk
MI adherent	appreciation, downplayer, thanking
MI non-adherent	action-directive, offers/options commits, 3rd-party-talk

Table 8: Label clustering results for therapist's codes .

Behavioral Code	Clustered Similar Labels
Follow/Neutral	backchannel, no answer, non-verbal, yes answer, response acknowledgement, tag-question, repeat-phrase, backchannel in question form
Change Talk (positive)	quotation, declarative, yes-no-question, offers/options commits, statement-opinion, declarative wh-question, rhetorical-questions, 3rd-party-talk, yes-no-question
Sustain Talk (negative)	statement-non-opinion, collaborative completion, hedge, action-directive, other answers, dispreferred answers, declarative yes-no-question, affirmative non-yes answers

Table 9: Label clustering results for patient's codes .