Interspeech 2022
18-22 September 2022, Incheon, Korea

Towards Automated Counselling Decision-Making: Remarks on Therapist Action Forecasting on the AnnoMI Dataset

Zixiu Wu[1,2]*, Rim Helaoui*[1]*, Diego Reforgiato Recupero*[2]*, Daniele Riboni*[2]

[1]Philips Research, Eindhoven, The Netherlands
[2]University of Cagliari, Cagliari, Italy

{zixiu.wu, rim.helaoui}@philips.com, {diego.reforgiato, riboni}@unica.it

Abstract

Substantial progress has been made in recent years on natural language processing approaches to counselling conversation analysis. However, few studies have investigated therapist action forecasting, which aims to suggest dialogue actions that the therapist can take in the next turn, partly due to generally limited access to counselling dialogue data resulting from privacy-related constraints. In this work, we leverage a recently released public dataset of therapy conversations and experiment with a range of natural language processing techniques to approach the task of therapist action forecasting with language models. We probe various factors that could impact model performance, including data augmentation, dialogue context length, incorporating therapist/client utterance labels in the input, and contrasting high- and low-quality counselling dialogues. With our findings, we hope to provide insights on this task and inspire future efforts in counselling dialogue analysis.

Index Terms: Counselling, Therapy, Motivational Interviewing, Dialogue, Natural Language Processing

1. Introduction

Dialogue-related natural language processing (NLP) research has seen significant development recently, driven by increasingly powerful language models (LMs). In domains with counselling-like elements such as peer support dialogue, progress has been made ranging from empathy detection ([1, 2, 3], *inter alia*) to empathetic response generation (e.g., [4]). NLP for counselling dialogue analysis, however, has not been developed to the same extent, mostly due to privacy constraints in using real therapy conversation data. Nevertheless, recent works have looked into topics such as examining therapist strategies [5] and providing real-time counsellor evaluation [6].

In this work, we focus on a counselling style called Motivational Interviewing (MI) [7, 8], which is widely used in healthcare to help the client achieve a target behaviour change such as smoking cessation and improve their health. The core of MI is to evoke the client's own motivation to change. The therapist in high-quality MI centres on the client and expresses empathy, while they mainly give instructions and suggestions in low-quality MI [7]. To monitor counselling quality and train therapists, coding systems such as the Motivational Interviewing Skill Codes (MISC) [9] have been developed to classify therapist and client utterances into a set of behaviour codes.

For MI-related NLP, researchers have explored applications such as therapist empathy modelling [10, 11, 12, 13, 14], automatic coding of therapist/client utterances [15, 16, 17, 18], and reflection generation [19], with methods ranging from classical machine learning with linguistic features to advanced deep learning models. In particular, [18] established two complementary MI-based tasks — **code categorisation** and **code forecasting**. Given a dialogue history (context), the former assigns a MISC code to the latest therapist/client utterance, while the latter forecasts the code for the next utterance. Reliable therapist code forecasting can assist the therapist in an ongoing session and be conducive to automated counselling decision-making.

We investigate the task of **therapist action forecasting**[1], which is slightly different from **code forecasting** [18] in that 1) MISC-inspired therapist dialogue actions[2] instead of actual MISC codes are forecasted, the former defined through regrouping and simplification of the latter based on counselling principles, and 2) we focus on forecasting only therapist actions and not client behaviours. To the best of our knowledge, no similar efforts have been made since [18] where GRUs were used for code forecasting, while NLP models have evolved significantly since then. Recently, the release of AnnoMI [20], an MI dialogue dataset expert-annotated at utterance level with therapist actions and client behaviours, has enabled further research into MI dialogue analysis, hence we leverage AnnoMI for this task.

Utilising LMs, we probe a range of modelling choices and descriptively analyse their effects on forecasting performance, including 1) varying dialogue history length, 2) using data augmentation to expand training data, 3) inserting therapist action and/or client behaviour labels in the context, and 4) contrasting therapist actions in high- and low-quality MI.

Our experiments show that the baseline of using the original dialogue contexts without special processing obtains the best performance, with minor contribution from context length. The modelling choices explored are mostly not conducive to better performance, which we posit is due to the noise introduced in the process. The best-performing model does not produce highly accurate forecasting if only the top-1 result is used, which reflects the latitude of a therapist in responding to the client. We make our code publicly available[3] for benchmarking and to inspire future efforts in counselling dialogue analysis.

2. Methodology

2.1. Problem Statement

At time step t in a counselling dialogue, we take the most recent N turns (one utterance per turn) as the dialogue history, namely $H_t^N = \{u_{t-N}, \cdots, u_{t-1}\}$ where u_{t-1} is a client utterance, and the goal is to forecast the therapist action label y_t^T of the immediate therapist response u_t, given H_t^N as the input. $y_t^T \in Y^T$ where Y^T is a predefined set of therapist action labels.

[1]We use "forecasting" to refer to generating the next-turn action given the dialogue history, while "classification"/"prediction" refers to determining the current-turn action given only the current turn.

[2]Referred to as "(main) therapist behaviours" in [20].

[3]https://github.com/zixiu-alex-wu/InterspeechMIPublic

10.21437/Interspeech.2022-506

Table 1: *AnnoMI overview*

	High-Quality MI	Low-Quality MI
#Dialogues	110 (82.7%)	23 (17.3%)
#Utterances	8839 (91.1%)	860 (8.9%)

Table 2: *Utterance label distribution in AnnoMI*

Attribute	Labels	Prop. in High-Qlt. MI Conv.	Prop. in Low-Qlt. MI Conv.
Therapist Action	**Reflection**	28%	7%
	Question	28%	32%
	Input	11%	33%
	Other	33%	28%
Client Talk Type	**Change**	25%	17%
	Neutral	64%	68%
	Sustain	11%	15%

2.2. Counselling Dialogue Data

We use the AnnoMI dataset [20], which consists of 133 conversations[4] covering a wide range of topics such as "reducing alcohol use" and "smoking cessation". Each conversation was transcribed from an expert-created demonstration video and then labelled as demonstrating high-/low-quality MI based on the video title and description. The 110 dialogues showcasing high-quality MI with over 8.8K utterances in total (Table 1) are used to form the training/validation/test sets for this task, as the goal is to emulate dialogue actions that **good** therapists would take. We note, however, that we do utilise low-quality MI dialogues as part of the training data in some setups (§2.4.4).

For $\{y_t^T\}$, we use the annotations of AnnoMI. Each therapist utterance is annotated as one of the four dialogue actions: $Y^T = \{$**Reflection**, **Question**, **Input**, **Other**$\}$. **Reflection** (reflective listening) conveys the effort of listening and trying to understand the client, **Question** includes open- and closed questions, **Input** consists of offering information and advice, etc., while **Other** applies when no other behaviour is present, e.g., greetings and facilitators like "Mm-hmm". Notably, asking, informing and listening (esp. reflective listening) are three basic but important communication skills in MI [8], and thus **Question**, **Input** and **Reflection** are correspondingly established as dialogue actions in AnnoMI. In practice, Y^T regroups and simplifies MISC codes: **Reflection** and **Question** are also MISC codes, while **Input** incorporates codes like "Giving Information" and "Advise", and, similarly, **Other** encompasses codes like "Facilitate" and "Filler". The proportions of utterances with those labels are shown in Table 2.

Each client utterance in AnnoMI is annotated with a "talk type" (behaviour): **Change**, **Neutral**, or **Sustain**, which indicates leaning towards (**Change**), moving away from (**Sustain**), or showing no inclination (**Neutral**) w.r.t. the target behaviour change. We leverage this attribute in some setups (§2.4.3).

2.3. General Input Format

To distinguish between the utterances in a dialogue history, we insert interlocutor labels and utterance separators as plain text. For example, a 3-turn context H_t^3 shown in Table 3 is converted into a single sequence as input to the model as follows:

[4]The therapist-client pairing of each dialogue may not be unique.

Table 3: *A 3-turn context from a high-quality MI conversation*

Utt.	Role	Oracle	Text
u_{t-3}	**Client**	**Neutral**	I guess I'm not paying attention to it.
u_{t-2}	**Therapist**	**Reflection**	Yeah. There's certainly been– There's no problems and you, as you said, only have had it for a short time.
u_{t-1}	**Client**	**Neutral**	Mm-hmm.

"⟨client⟩I guess I'm not paying attention to it.|⟨therapist⟩Yeah. There's certainly been– There's no problems and you, as you said, only have had it for a short time.|⟨client⟩Mm-hmm."

2.4. Modelling Choices

Fine-tuning LMs on AnnoMI for the task, we explore several modelling choices. While recent work (e.g., [21]) has shown the superior few-shot performance of large-scale LMs, we leave the probing of few-shot learning for this task to future work.

2.4.1. Dialogue History Length

More knowledge of the dialogue exchanges so far may lead to better therapist action suggestion. Thus, we vary the dialogue history length (i.e., number of utterances) to probe its effects. For every other modelling choice in §2.4, we combine it with different context lengths for deeper insights.

2.4.2. Data Augmentation

The relatively small scale of AnnoMI motivates the use of data augmentation. We opt for paraphrasing as the means of augmentation, in order to minimize modification of the semantics of the dialogue history while increasing syntactic diversity [22].

2.4.3. Inserting Utterance Labels in Context

As utterance-level labels offer MI-relevant details, incorporating them in the context may impact task performance. We explore three options of inserting utterance labels as plain text:

- **Therapist Only:** prepending the label of each therapist utterance to the utterance as plain text.
- **Client Only:** prepending the label of each client utterance to the utterance as plain text.
- **Therapist & Client:** prepending the label of each therapist and client utterance to the utterance as plain text.

Since ground-truth utterance labels are not available during inference time, we experiment with three label sources:

- **Oracle:** using ground-truth utterance labels.
- **Predicted:** training a current-turn utterance label classifier (only current-turn utterance as input) and using its predicted label for each utterance in the dialogue history.
- **Random:** using randomly sampled utterance labels, for comparison with **Oracle** and **Predicted**.

As an example, the context H_t^3 in Table 3 is converted into the following using **Therapist & Client** and **Oracle**: *"⟨client⟩~⟨neutral⟩I guess I'm not paying attention to it.|⟨therapist⟩~⟨reflection⟩Yeah. There's certainly been– There's no problems and you, as you said, only have had it for a short time.|⟨client⟩~⟨neutral⟩Mm-hmm."*

2.4.4. Contrasting High- & Low-Quality MI

Inspired by recent work using plain-text control codes to influence LM output (e.g., [23]), we probe contrasting high- & low-quality MI with plain-text MI quality labels, since low-quality MI as negative examples may improve decision boundaries.

Specifically, we prepend the MI quality label of the conversation from which a context H_t^N is taken, while the other parts of the input remain unchanged. For the high-quality MI H_t^3 shown in Table 3, the input becomes: *"[high][SEP]⟨client⟩I guess I'm not paying attention to it.|⟨therapist⟩Yeah. There's certainly been– There's no problems and you, as you said, only have had it for a short time.|⟨client⟩Mm-hmm."*, where *[SEP]* is a model-specific separator for a pair of texts. The models are trained and evaluated on contrasting high- & low-quality MI dialogues and then tested on high-quality MI conversations only.

Due to the imbalance between high- and low-quality MI dialogue volumes (Table 1), we explore two variants of contrast:

- **Unbalanced:** using the original unbalanced high- and low-quality MI dialogue contexts.
- **Aug-Balanced:** using augmented low-quality MI contexts as described in §2.4.2 to achieve balance.

3. Experiments

3.1. Predicted Utterance Labels to Insert in Context

To obtain **Predicted** utterance labels to insert in the dialogue history (§2.4.3), we train two separate classifiers for therapist and client utterances. With 5-fold cross validation, a **Predicted** label is assigned to each therapist and client utterance in `AnnoMI`, which we then incorporate into the input accordingly.

Fine-tuning `roberta-base` [24], we obtain a macro F1 (unweighted mean of per-class F1 scores) of 0.78 and 0.53 for therapist and client utterance classification, respectively. The lower performance of the latter can be attributed to the lower inter-annotator agreement of 0.47 (moderate) — as measured by Fleiss' kappa — on client talk type annotations compared to the substantial agreement (0.74) on therapist action annotations [20], which likely led to more noisy ground-truth labels.

3.2. Implementation Details

Considering the relatively small scale of `AnnoMI`, 5-fold cross validation is used in all our experiments, and the dialogues in the training, validation and test sets are mutually exclusive.

We consider 6 context length options: $\{1, 3, 5, 7, 9, max\}$, using the most recent 1/3/5/7/9 turns or using as much context as possible (max). Where applicable, we left-truncate the input to keep the 512 tokens representing the most recent context.

Based on a preliminary study, we use an off-the-shelf Pegasus [25]-based neural paraphraser[5] to generate syntactically diverse utterance paraphrases and minimize semantic change, where applicable. Specifically, we generate 10 paraphrases $\{\hat{u}_i^m\}_{m=1}^{10}$ for each utterance $u_i \in$ `AnnoMI`, and for each original dialogue history $H_t^N = \{\ldots, u_i, \ldots\}$ we obtain 5 augmentations $\{\tilde{H}_t^{N,o}\}_{o=1}^{5}$ where $\tilde{H}_t^{N,o} = \{\ldots, \tilde{u}_i^o, \ldots\}$ and $\tilde{u}_i^o$ is randomly sampled from $\{u_i\} \cup \{\hat{u}_i^m\}_{m=1}^{10}$. Thus, we use only $\{\tilde{H}_t^{N,o}\}_{o=1}^{5}$ during training to effectively train on 5x amount of data, while keeping the validation and test sets unchanged.

All our models are based on the HuggingFace [26] implementation of `roberta-base`[6], using an AdamW [27] optimiser with linear learning rate decay from an initial 2e-5. The batch size is 8, and the maximum input length is 512 tokens.

[5] https://huggingface.co/tuner007/pegasus_paraphrase

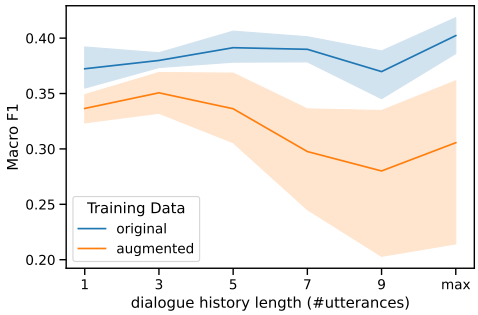

Figure 1: *Impact of context length & data augmentation*

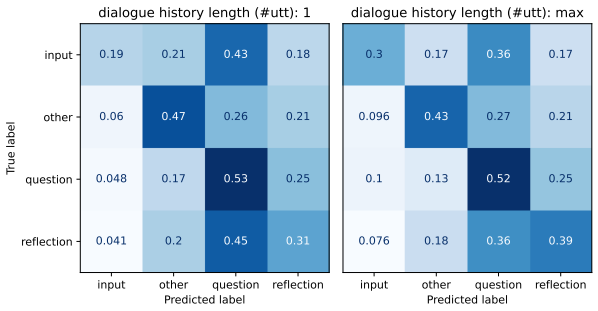

Figure 2: *Confusion matrices of* 1- *&* max*-utterance baselines*

3.3. Results & Analysis

With 5-fold cross validation, we obtain 5 test-set scores from the 5 splits for each training setup. Therefore, in Figures 1, 3 and 4, we show the mean score (line) with a 95% confidence interval (error band) based on the 5 test-set scores of each setup, which offers insights on cross-split performance variation. Unless otherwise specified, we use macro F1 as the metric, following [18].

3.3.1. Dialogue History Length & Data Augmentation

Training on unaugmented high-quality MI dialogues, the performance improves with longer contexts and reaches 0.39 macro F1 under the 5-utterance context setting (Figure 1). Afterwards, it steadily decreases as the context grows further but rebounds to 0.4 when using maximum history. Overall, the performance does not vary substantially across the splits.

Figure 2 shows the confusion matrices of the 1- and max-utterance models, where one observes that both models forecast **Question** most correctly, followed by **Other**, **Reflection** and **Input**. In particular, increasing the context length from 1 to max benefits the forecasting of **Input** and **Reflection** the most.

The best score of 0.4 macro F1 is insufficient for real-world deployment, echoing [18] where the best code forecasting score was 0.31, though the results are not directly comparable since theirs were based on an undisclosed counselling dialogue dataset annotated differently. The low score may be linked to the latitude of therapists in counselling, as sometimes there are multiple good actions to take. Just as the confusion matrices

[6] We also explored `roberta-large` which had consistently lower scores and larger cross-split performance variation in cross validation.

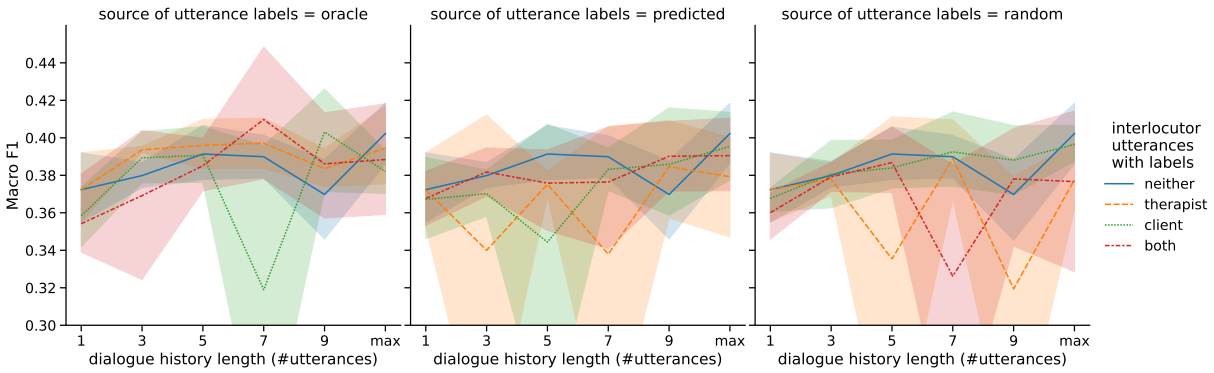

Figure 3: *Impact of inserting therapist/client utterance labels in dialogue history on model performance*

show the model uncertainty deciding between **Reflection** and **Question**, the therapist may reflect or pose a question after, for example, the client explains their personal circumstances.

Training on augmented context yields consistently lower performance and considerable cross-split performance variation. As the context grows longer — thus more paraphrases used in the input — the performance generally worsens and larger cross-split performance variation occurs. We hypothesise that therapist action forecasting is sensitive to conversation semantics, and that the performance is therefore negatively impacted by the altered semantics in some paraphrases caused by hallucination [28], a well-known problem of neural LMs where the generated text is unfaithful to the source input.

3.3.2. *Inserting Utterance Labels in Context*

Figure 3 shows the model performances when therapist and/or client utterance labels are incorporated in the dialogue history, where the label sources are **oracle**, **predicted**, or **random**.

Overall, using only **oracle** therapist utterance labels slightly ($\leq$ 0.01 macro F1) outperforms the label-less baseline in most settings, while using **oracle** labels for client utterances only and for both therapist and client utterances shows mixed results and larger cross-split performance variation. This difference likely points to the closer alignment between therapist utterances and their **oracle** labels as evidenced by the higher inter-annotator agreement, which enables useful additional training signals.

Using **predicted** and **random** utterance labels mostly underperforms the baseline and shows larger cross-split performance variation, with **random** suffering slightly more. While **random** introduces considerable noise in the input which unsurprisingly harms performance, the underperformance of **predicted** is unexpected, especially considering the relatively good performance of therapist utterance classification (§3.1). One possible explanation is that the label-less baseline already understands the context relatively well without slightly noisy predicted utterance labels, especially in setups with longer contexts where incorrect predicted labels are more likely to occur.

3.3.3. *Contrasting High- & Low-Quality MI*

The baseline trained only on high-quality MI dialogues mostly surpasses the models trained on mixed-quality MI conversations (Figure 4). Between the latter, **Aug-Balanced** generally yields lower scores, echoing the finding of §3.3.1 that training on paraphrased dialogues harms performance. One plausible reason for **Unbalanced** underperforming the baseline is that the therapist

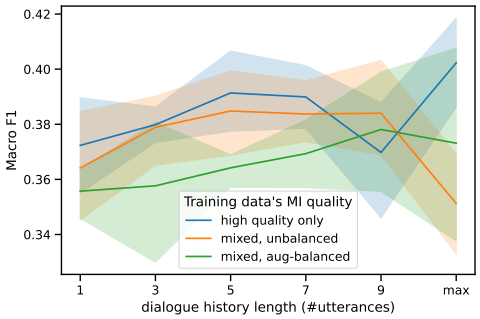

Figure 4: *Impact of contrasting high- & low-quality MI*

actions in the low-quality MI dialogues may not adequately represent the "mistakes" of the baseline, and hence the contrast is not effective enough to improve the decision boundaries, while the low-quality MI dialogues at the same time introduce more uncertainty into the ground truth action labels during training.

4. Discussion & Conclusion

We experimented with various modelling choices for therapist action forecasting, including dialogue history length, data augmentation, inserting utterance labels into the dialogue history, and contrasting high- and low-quality MI dialogues. Generally, the baseline using plain dialogue context without particular NLP techniques achieves the best results, with relatively minor impact from context length. The techniques explored in this work prove to mostly introduce noise and hurt performance.

Overall, the strong baseline is not an ideal forecaster if only the top-1 forecast is used. Since the ground-truth labels are well-defined and annotated with a substantial agreement, a likely explanation for the low performance is that it is linked to the latitude/flexibility of therapists in their response. Therefore, future work could investigate formulating this task probabilistically. Also worth exploring are 1) suggesting what action(s) the therapist should **not** take next and 2) detecting worsening of counselling quality in real-time, where the contrast between high- and low-quality MI dialogues is likely more beneficial.

5. Acknowledgements

This work has been funded by the EC in the H2020 Marie Skłodowska-Curie PhilHumans project, contract no. 812882.

6. References

[1] A. Sharma, A. Miner, D. Atkins, and T. Althoff, "A computational approach to understanding empathy expressed in text-based mental health support," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 5263–5276.

[2] N. Zhou and D. Jurgens, "Condolence and empathy in online communities," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[3] M. Hosseini and C. Caragea, "It takes two to empathize: One to seek and one to provide," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 14, 2021, pp. 13 018–13 026.

[4] A. Sharma, I. W. Lin, A. S. Miner, D. C. Atkins, and T. Althoff, "Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach," in *Proceedings of the Web Conference 2021*, 2021, pp. 194–205.

[5] J. Zhang and C. Danescu-Niculescu-Mizil, "Balancing objectives in counseling conversations: Advancing forwards or looking backwards," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5276–5289.

[6] A. Li, J. Ma, L. Ma, P. Fang, H. He, and Z. Lan, "Towards automated real-time evaluation in text-based counseling," *arXiv preprint arXiv:2203.03442*, 2022.

[7] W. R. Miller and S. Rollnick, *Motivational interviewing: Helping people change*. Guilford press, 2012.

[8] S. Rollnick, W. R. Miller, and C. Butler, *Motivational interviewing in health care: helping patients change behavior*. Guilford Press, 2008.

[9] W. R. Miller, T. B. Moyers, D. Ernst, and P. Amrhein, "Manual for the motivational interviewing skill code (misc)," *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*, 2003.

[10] B. Xiao, D. Can, P. G. Georgiou, D. Atkins, and S. S. Narayanan, "Analyzing the language of therapist empathy in motivational interview based psychotherapy," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2012, pp. 1–4.

[11] J. Gibson, N. Malandrakis, F. Romero, D. C. Atkins, and S. S. Narayanan, "Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms," in *Sixteenth annual conference of the international speech communication association*, 2015.

[12] J. Gibson, D. Can, B. Xiao, Z. E. Imel, D. C. Atkins, P. Georgiou, and S. S. Narayanan, "A Deep Learning Approach to Modeling Empathy in Addiction Counseling," in *Proc. Interspeech 2016*, 2016, pp. 1447–1451.

[13] Z. Wu, R. Helaoui, V. Kumar, D. Reforgiato Recupero, and D. Riboni, "Towards detecting need for empathetic response in motivational interviewing," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 497–502.

[14] Z. Wu, R. Helaoui, D. R. Recupero, and D. Riboni, "Towards low-resource real-time assessment of empathy in counselling," in *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, 2021, pp. 204–216.

[15] D. C. Atkins, M. Steyvers, Z. E. Imel, and P. Smyth, "Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification," *Implementation Science*, vol. 9, no. 1, pp. 1–11, 2014.

[16] B. Xiao, D. Can, J. Gibson, Z. E. Imel, D. C. Atkins, P. G. Georgiou, and S. S. Narayanan, "Behavioral coding of therapist language in addiction counseling using recurrent neural networks." in *Interspeech*, 2016, pp. 908–912.

[17] J. Gibson, D. Atkins, T. Creed, Z. Imel, P. Georgiou, and S. Narayanan, "Multi-label multi-task deep learning for behavioral coding," *IEEE Transactions on Affective Computing*, 2019.

[18] J. Cao, M. Tanana, Z. Imel, E. Poitras, D. Atkins, and V. Srikumar, "Observing dialogue in therapy: Categorizing and forecasting behavioral codes," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5599–5611.

[19] S. Shen and C. Welch, "Counseling-style reflection generation using generative pretrained transformers with augmented context," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2020.

[20] Z. Wu, S. Balloccu, V. Kumar, R. Helaoui, E. Reiter, D. R. Recupero, and D. Riboni, "Anno-mi: A dataset of expert-annotated counselling dialogues," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6177–6181.

[21] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[22] B. Li, Y. Hou, and W. Che, "Data augmentation approaches in natural language processing: A survey," *arXiv preprint arXiv:2110.01852*, 2021.

[23] H. Rashkin, D. Reitter, G. S. Tomar, and D. Das, "Increasing faithfulness in knowledge-grounded dialogue with controllable features," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 704–718.

[24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[25] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 328–11 339.

[26] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[28] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *arXiv preprint arXiv:2202.03629*, 2022.