ANNO-MI: A DATASET OF EXPERT-ANNOTATED COUNSELLING DIALOGUES

Zixiu Wu^{*†12}*, Simone Balloccu*^{*3}*, Vivek Kumar*^{21}*,*
Rim Helaoui^{1}*, Ehud Reiter*^{3}*, Diego Reforgiato Recupero*^{2}*, Daniele Riboni*^{2}

^{1}Philips Research, Eindhoven, Netherlands
^{2}Department of Mathematics & Computer Science, University of Cagliari, Cagliari, Italy
^{3}Department of Computing Science, University of Aberdeen, Aberdeen, UK

†zixiu.wu@philips.com

ABSTRACT

Research on natural language processing for counselling dialogue analysis has seen substantial development in recent years, but access to this area remains extremely limited due to the lack of publicly available expert-annotated therapy conversations. In this work, we introduce AnnoMI, the first publicly and freely accessible dataset of professionally transcribed and expert-annotated therapy dialogues. It consists of 133 conversations that demonstrate high- and low-quality motivational interviewing (MI), an effective counselling technique, and the annotations by domain experts cover key MI attributes. We detail the data collection process including dialogue selection, transcription and annotation. We also present analyses of AnnoMI and discuss its potential applications.

Index Terms— Counselling, Motivational Interviewing, Dialogue, Natural Language Processing, Dataset

1. INTRODUCTION

Behaviour change such as smoking cessation could greatly improve patient health, but it can be challenging for counsellors to promote it to patients [1]. Therefore, a therapeutic technique called Motivational Interviewing (MI) has been developed for eliciting the motivation to change from the client[1] themselves [2]. Correspondingly, coding systems such as MISC [3] and MITI [4] are widely used to capture therapist- & client-related MI codes and aspects.

Aside from speech features ([5, 6, 7], *inter alia*), linguistic features have also been utilised to analyse MI therapeutic language with statistical models. Can et al. [8] proposed the first computational model for identifying *Reflection*, an important therapeutic skill. Classical machine learning [9, 10, 11] and deep learning methods [12, 13, 14, 15] have both been leveraged to predict MI codes and aspects such as therapist empathy. More recently, Pérez-Rosas et al. [16] released a dataset of high- & low-quality-MI[2] dialogues taken from online video-sharing platforms and analysed the linguistic aspects that distinguish between MI-adherent and non-adherent therapy. On the other hand, Wu et al. [17, 18] explored predicting therapist empathy in low-resource settings.

Despite its progress, MI-related natural language processing (NLP) has been limited by the lack of publicly available MI conversations due to privacy constraints. Most work has been based on undisclosed corpora of MI dialogues and annotations, making it difficult to reproduce and build on previous findings. To the best of our knowledge, the only publicly and freely accessible MI corpus is from [16], based on YouTube/Vimeo videos transcribed with automatic speech recognition (ASR). However, the transcripts contain considerable ASR noise and incorrect interlocutor labels (i.e. client utterances labelled as therapist utterances, and vice versa) that hinder understanding. [16] also analysed annotations of two key MI codes — *Reflection* and *Question*, but those annotations are unavailable in the dataset at the time of writing.

To address the lack of publicly available expert-annotated MI dialogues and improve access to MI-related NLP research, we present AnnoMI, a dataset[3] of 133 high- and low-quality MI conversations that 1) were professionally transcribed from MI demonstration videos on video-sharing platforms, 2) were obtained through explicit consent from the video owners that permits dataset creation and release to the public and use for research purposes, and 3) are annotated by experienced MI practitioners based on a scheme covering key MI aspects. We describe the data collection process in Section 2, including obtaining consent, transcription and annotation. Section 3 presents dataset analyses and discusses potential applications of AnnoMI, and Section 4 concludes the paper.

*Equal contribution

†Knowledge contributor & corresponding author

This work has been funded by the EC in the H2020 Marie Skłodowska-Curie PhilHumans project, contract no. 812882. We would also like to thank Dr. Mark Aloia for his guidance and support.

[1] A client of behaviour change therapy may not have an illness, so we use the term "client" instead of "patient" throughout the remainder of this work.

[2] In this work, "MI-adherent" is used as a synonym of "high-quality" and similarly "MI non-adherent" is identical to "low-quality". These terms are not related to video quality or transcription quality.

[3] Available at https://github.com/uccollab/annomi under the Public Domain license.

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

Table 1. Dataset overview

2. BUILDING COUNSELLING DATASET

Due to the lack of publicly available dialogue datasets of real-world motivational interviewing as well as their privacy-related constraints, we opt for demonstrations of high- and low-quality MI on online video-sharing platforms, inspired by [16]. With explicit consent from the video owners, we had those demonstration videos professionally transcribed and then obtained annotations from experienced MI practitioners.

2.1. MI Demonstration Videos

As a trade-off between conversation authenticity and privacy-related constraints, we limit the sources of dialogues to online video-sharing platforms (YouTube & Vimeo). Based on [16] and exhaustive searching, we identified 346 demonstration videos of high- and low-quality MI with keywords such as "effective MI"/"good MI" and "ineffective MI"/"bad MI". In terms of definitions of high/low-quality MI, we note that according to general counselling principles from the literature [2], the therapist in a high-quality session centres on the client and expresses empathy, while the counterpart in a low-quality session mostly provides instruction and suggestions.

We labelled the videos as high- or low-quality MI, based directly on the titles (e.g. "MI-Good example", "How NOT to do Motivational Interviewing") and/or descriptions and comments from the narrators (e.g. "This is ... video ... where I demonstrate how to use motivational interviewing ...") of the videos. Those MI quality labels are automatically validated, since the videos are all from professional therapists and established institutions for MI & positive behaviour change.

We contacted the video owners to seek their explicit consent[4] to the videos being used for dialogue analysis and dataset creation & release. Eventually, we were authorised to use 119[5] of the videos, containing 133 complete conversations in total (a video may contain multiple dialogues), with 110 conversations showcasing MI-adherence and 23 demonstrating non-adherence, as summarised in Table 1. Table 2 shows conversations excerpts of high- and low-quality MI.

The imbalance between high- and low-quality-MI conversations is due to 1) the relative lack of consent from low-quality-MI video owners and 2) the general lack of low-quality-MI demonstrations on video-sharing platforms, potentially because high-quality-MI videos as "positive examples" are considered more valuable and thus shared more.

High-Quality MI
T: Mm-hmm. So it's kind of surprising to you that something you've been doing and you've been doing more and more of it is actually pretty bad for you. ***C***: Oh, yes. I checked the box on your form when you asked if I use tobacco, I checked "No" because I never thought of myself as a smoker. ***T***: Mm-hmm. What do you kind of make of that now that you realize that you're actually a tobacco user and that you might actually be causing some pretty se-serious health effects.
Low-Quality MI
T: So you're gonna quit then? ***C***: Uh, maybe. ***T***: What do you mean, maybe? I just told you how bad it is for you. It's messing up your mouth, you're putting yourself at risk for all these other diseases. This is really important. You need to quit.

Table 2. Smoking cessation dialogue excerpts from high- & low-quality MI. ***T***: therapist; ***C***: client.

2.2. Transcription

To maximise transcript quality, we used a professional transcription service[6] to obtain the transcripts of the videos, thus obtaining faithfully transcribed utterances that are more understandable, in contrast to the transcripts of [16] obtained via automatic captioning which leads to transcription errors and incorrect therapist/client interlocutor labels. In addition, our approach removes noise such as narrations that do not form part of a dialogue and keeps conversation details that contextualise utterances, including "hmm" and "right" as well as "[laugh]" which indicates laughter by the interlocutor.

2.3. Expert Annotators & Workload Assignment

We opt for expert annotations on the transcripts, since annotation of therapeutic dialogue requires specialised training. 10 therapists found through the Motivational Interviewing Network of Trainers[7] , an international organisation of MI trainers and a widely recognised authority in MI, were recruited for the task. All the annotators were highly proficient in English and had prior experience in practising and coding MI.

Each annotator was randomly assigned 19 to 20 transcripts of varying lengths that totalled about 144 minutes in terms of the durations of the original videos. In order to investigate inter-annotator agreement (IAA), 7 common transcripts, totalling 45 minutes and with varying lengths & MI qualities, were annotated by all the annotators. The annotators were unaware that 7 of the transcripts annotated by them would be used to compute the IAA.

[4] The consent of the individuals in the videos was gathered together with that of the content owner where applicable.

[5] 42 of the 119 videos are overlapped with [16].

[6] `https://gotranscript.com/`

[7] `https://motivationalinterviewing.org/`

Attribute	Labels	Label Sub-Types
(Main) *Therapist* *Behaviour*	**Reflection**	Simple Reflection Complex Reflection
	Question	Open Question Closed Question
	Input	Information Advice Giving Options Negotiation/Goal-Setting
	Other	
Client Talk Type	**Change** **Neutral** **Sustain**	

Table 3. Utterance-level annotation scheme, every attribute annotated in the form of multi-choice. Label sub-types are for reference purposes only.

Topic	#Dialogues
Reducing alcohol consumption	28 (21.1%)
Smoking cessation	21 (15.8%)
Weight loss	9 (6.8%)
Taking medicine / Following medical procedure	9 (6.8%)
More exercise / Increasing activity	9 (6.8%)
Reducing drug use	8 (6.0%)
Reducing recidivism	7 (5.3%)
Other	48 (36.1%)

Table 4. Behaviour change topics in `AnnoMI`, with the number and percentage of conversations where a topic occurs. Note that a conversation can have more than one topic.

2.4. Annotation Scheme

Inspired by MI literature and suggestions from therapists, we study therapist behaviours and client talk types at utterance level (Table 3). The annotators annotate therapist behaviour for each therapist utterance and client talk type for each client utterance, and select only one label for each attribute. At conversation level, the annotators briefly describe the goal(s) of each conversation (e.g. "smoking cessation"), based on which we summarise the conversation topics in Table 4.

2.4.1. (Main) Therapist Behaviour

ASKING, INFORMING and LISTENING are three basic but important communication skills in MI that enable efficient and productive consultations [1]. Based on this observation and relevant MI coding rules, we establish **Question**, **Input**, and **Reflection** as three major therapist behaviours for analysing ASKING, INFORMING and LISTENING respectively. We also list some sub-types for each behaviour, but since a preliminary study showed low inter-annotator agreement on behaviour sub-type level, we only provide the list of sub-types to the annotators for reference purposes. In cases where a therapist utterance contains multiple behaviours such as a reflection followed by a question, the annotator is required to choose the **main behaviour**. We also add **Other** as a fourth behaviour that applies where no **Question**, **Input** or **Reflection** is manifested in the utterance.

Interlocutor	Utterance	Reflection
Client	At one time I was pretty much anti anything but marijuana	
Therapist 1	Marijuana was OK	Simple
Therapist 2	That's where you drew the line	Complex

Table 5. Simple & complex reflection. For reference purposes only — **Reflection** sub-types are not annotated.

ASKING is used to develop an understanding of the client and their situation. We thus designate **Question** as a therapist behaviour, and follow MISC in defining a question to be either *open* or *closed*, which is similar to open-ended/closed-ended questions but with some nuanced differences (e.g. "tell me more" is an open question).

INFORMING is the principal way of communicating knowledge to the client. Based on mainstream MI coding systems, we rephrase *Informing* as **Input** and use it as a therapist behaviour to cover a range of conveyed knowledge. Concretely, we name *advice*, *information*, *giving options* and *negotiation/goal-setting* as the sub-types of **Input**.

A main way of LISTENING is reflective listening, which is an important quality indicating listening, hearing and understanding the client. Thus, we include **Reflection** as a therapist behaviour and consider two sub-types: *simple* & *complex*, following MISC. Simple reflection conveys understanding of what the client has said and adds little extra meaning, e.g. through paraphrasing, while complex reflection shows a deeper understanding of the perspective of the client and adds substantial meaning to their words, e.g. through analogy and summary [1]. A pair of simple & complex reflections to the same client utterance are shown in Table 5.

Finally, we consider a therapist utterance to be of **Other** behaviour if no **Input**, **Reflection** or **Question** is involved, such as greetings and simple utterances like "Mhmm".

2.4.2. Client Talk Type

As pointed out in MI literature [1], clients are often ambivalent about positive behaviour change, and therefore a key goal of MI is for clients to convince themselves to change, if it is compatible with their personal values and aspirations. Such talks for change are referred to as "change talks", while "sustain talks" show resistance to change and an inclination to maintain the status quo. Finally, "neutral talks" do not signal leaning towards or away from change. Thus, we define **Change**, **Sustain** and **Neutral** as the three labels of the **Client**

Utterance	Talk Type
My doc told me I'm gonna lose my leg if I don't start checking my blood sugars	Change
I hate a night without a buzz	Sustain
Uh huh	Neutral

Table 6. Example Labelling for **Client Talk Type**.

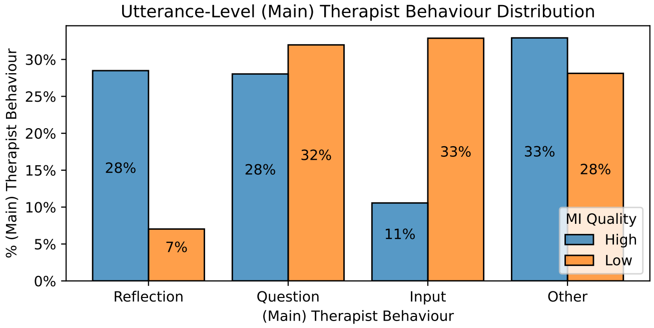

Fig. 1. Distribution of utterance-level therapist behaviours in high- & low-quality MI.

Talk Type attribute. Table 6 shows some examples.

2.5. Inter-Annotator Agreement

We use Fleiss' kappa to measure the inter-annotator agreement (IAA) between the 10 annotators over their annotations on the 7 transcripts for IAA calculation, recording 0.74 (substantial agreement) and 0.47 (moderate agreement) for **(Main) Therapist Behaviour** and **Client Talk Type**, respectively. Therefore, we consider those attributes to be **predictable** and conduct dataset analysis on that basis.

3. DATASET ANALYSES & POTENTIAL APPLICATIONS

We analyse the distributions of utterance attributes in high- & low-quality MI. We note that while there are clear correlations between utterance attribute distribution and MI quality, they do not necessarily point to causation, especially given the relatively low amount of data and potential sampling bias.

As shown in Figure 1, the most pronounced difference between therapist behaviours in high- and low-quality MI is the use of **Reflection** and **Input**. The average MI-adherent therapist uses reflective listening 28% of the time while the proportion is only 7% in low-quality MI, which may be because high-quality MI requires trying to understand the perspective of the client and conveying it to them. In contrast, the average MI-non-adherent therapist gives **Input** in 33% of their utterances, while the MI-adherent counterpart does so only 11% of the time, which, when considered along with **Reflection**, is in line with the general observation [1] that high-quality MI is focused more on understanding the client and less on talking from their own viewpoint. On the other hand, the use of **Question** and **Other** is more independent of MI quality.

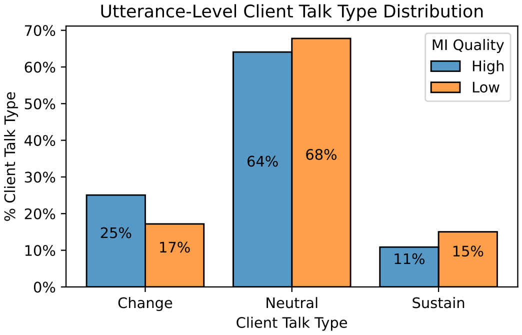

Fig. 2. Distribution of utterance-level client talk types in high- & low-quality MI.

From the talk type distribution in Figure 2, it is clear that **Change** talk emerges more in MI-adherent therapy — 25% vs. 17%, while **Sustain** talk is more present in low-quality counselling — 11% vs. 15%. However, these differences are less noticeable than those observed in **Reflection** and **Input**, potentially because 1) some clients in low-quality counselling might use unenthusiastic **Change**-talk-like language such as "Yeah, maybe" merely to exit the consultation as quickly as possible and 2) some clients in high-quality counselling are simply less willing to change and their therapists still respect that unwillingness, as advised in MI guidelines. Most client utterances are **Neutral** talks, which can be attributed to the large amount of simple utterances like "Mhmm".

In terms of potential applications, `AnnoMI` will facilitate future research thanks to its professional transcription and expert annotations. Basic applications include predicting current-turn therapist behaviour / client talk type and forecasting these two properties in the next turn, both of which can be leveraged to assess, coach and guide counsellors. Furthermore, pre-trained language models can be fine-tuned on `AnnoMI` to imitate client/therapist language and, for instance, generate potential next-turn therapist utterances to assist the human counsellor in a live session.

4. CONCLUSION

We introduce a new dataset of professionally transcribed and expert-annotated counselling dialogues demonstrating high- and low-quality motivational interviewing. We also analyse the distributions of utterance attributes and discuss potential applications of the dataset. Future work will involve the discussed potential applications and also cross-domain experiments, e.g. bridging general and therapeutic dialogue. We hope this dataset will improve access to and stimulate research in counselling-related natural language processing.

5. REFERENCES

[1] Stephen Rollnick, William R Miller, and Christopher Butler, *Motivational interviewing in health care: helping patients change behavior*, Guilford Press, 2008.

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

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

[4] Theresa B Moyers, Lauren N Rowell, Jennifer K Manuel, Denise Ernst, and Jon M Houck, "The motivational interviewing treatment integrity code (miti 4): rationale, preliminary reliability and validity," *Journal of substance abuse treatment*, vol. 65, pp. 36–42, 2016.

[5] Bo Xiao, Daniel Bone, Maarten Van Segbroeck, Zac E Imel, David C Atkins, Panayiotis G Georgiou, and Shrikanth S Narayanan, "Modeling therapist empathy through prosody in drug addiction counseling," in *Fifteenth annual conference of the international speech communication association*, 2014.

[6] Bo Xiao, Zac E Imel, David C Atkins, Panayiotis G Georgiou, and Shrikanth S Narayanan, "Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[7] Karan Singla, Zhuohao Chen, David Atkins, and Shrikanth Narayanan, "Towards end-2-end learning for predicting behavior codes from spoken utterances in psychotherapy conversations," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3797–3803.

[8] Doğan Can, Panayiotis G Georgiou, David C Atkins, and Shrikanth S Narayanan, "A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[9] Bo Xiao, Dogan Can, Panayiotis G Georgiou, David Atkins, and Shrikanth 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.

[10] David C Atkins, Mark Steyvers, Zac E Imel, and Padhraic 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.

[11] James Gibson, Nikolaos Malandrakis, Francisco Romero, David C Atkins, and Shrikanth 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] James Gibson, Doğan Can, Bo Xiao, Zac E. Imel, David C. Atkins, Panayiotis Georgiou, and Shrikanth S. Narayanan, "A Deep Learning Approach to Modeling Empathy in Addiction Counseling," in *Proc. Interspeech 2016*, 2016, pp. 1447–1451.

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

[14] James Gibson, David Atkins, Torrey Creed, Zac Imel, Panayiotis Georgiou, and Shrikanth Narayanan, "Multilabel multi-task deep learning for behavioral coding," *IEEE Transactions on Affective Computing*, 2019.

[15] Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek 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.

[16] Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea, "What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 926–935.

[17] Zixiu Wu, Rim Helaoui, Vivek Kumar, Diego Reforgiato Recupero, and Daniele 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.

[18] Zixiu Wu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele 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.