Incorporating Structural Information in Text Classification for Motivational Interviewing

Hosein Rezaei
Department of Computer Science
University of York
York, UK
hosein.rezaei@york.ac.uk

Tommy Yuan
Department of Computer Science
University of York
York, UK
tommy.yuan@york.ac.uk

Tarique Anwar
Department of Computer Science
University of York
York, UK
tarique.anwar@york.ac.uk

***Abstract*—Speech act classification is a crucial task in the application of dialogue systems for mental healthcare. Motivational Interviewing (MI) in particular, requires at least two speech act classifiers for classifying the utterances of both client and clinician. State-of-the-art MI classifiers, despite their good performance, aren't still accurate enough for being applied in a dialogue system. In order to improve their performance, we propose the use of different kinds of graphs such as Abstract Meaning Representation (AMR) or dependency graphs, either as direct input to graph neural networks or as a new attention layer in transformers. The impact of incorporating the structural information distilled in these graphs on classification will be reported.**

***Index Terms*—Text Classification, Motivational Interviewing, Abstract Meaning Representation, Dependency Trees,**

I. Introduction

Motivational Interviewing is one of mental healthcare therapeutic methodologies which is designed primarily around the concept of dialogue between the clinician and the client [1]. During an MI dialogue they try collaboratively to find and activate the personal reasons for changing the unexpected behaviours [2]. The core idea of MI is addressing ambivalence that can be manifested by "change talk" in which the client expresses their motivation for change, and by "sustain talk" wherein they show signs of resistance. The counselor on the other hand tries to promote the former by using MI consistent responses and addressing the latter.

The fact that MI is centered around dialogue, and even quality of MI is measured by some metrics defined over dialogue, has made it an interesting area for applying AI and dialogue systems. Recently, Large Language Models (LLMs) have proved very useful in generating meaningful and relevant responses, however, they suffer from lack of planning and the ability to follow a strategy toward achieving a goal [3]. Thus, an appealing line of research is enhancing dialogue capabilities of LLMs in order to make them suitable for performing MI dialogues.

One of the essential requirements of an MI dialogue system is detecting MITI codes in utterances of clinician [4], whereby we can determine their speech act (what act the clinician has tried to make by uttering a specific sentence). These are a set of 15 labels that help an observer (supervisor/evaluator) evaluate the quality of dialogue. For example, labels 1 to 4 in the (Table I) are considered "MI-inconsistent" and an LLM should avoid generating responses which might be categorised as any of them. On the other hand, labels 5 to 9 are known as MI-consistent, and the dialogue system is expected to generate more utterances with such speech acts.

TABLE I
List of MITI labels, used to determine speech act of clinician in MI

#	*Label*	*MI Guidance*
1	Confront	Inconsistent
2	Warn	Inconsistent
3	Direct	Inconsistent
4	Advice without Permission	Inconsistent
5	Advice with Permission	Consistent
6	Affirm	Consistent
7	Support	Consistent
8	Emphasize Autonomy	Consistent
9	Open Question	Encouraged
10	Closed Question	Discouraged
11	Simple Reflection	Encouraged
12	Complex Reflection	Encouraged*
13	Give Information	-
14	Self-Disclose	-
15	Other	-

*In MI, complex reflections are encouraged more over simple ones.

Another requirement of such a system is detecting "Change-Talk"s and "Sustain-Talk"s in utterances of client, as the eventual goal of MI dialogue is to maximize the number of former and to minimize that of the latter. Obviously, not every utterance by client is either of them and the rest should be categorized as "Neutral". Detecting these three labels in the sentences expressed by client, not only acts as a metric to determine if the goal of MI is achieved or not, but also is a key factor in deciding what to react to the client [5]. If a Change-Talk is detected, LLM should try to encourage and promote that. Conversely, if a Sustain-Talk is detected, LLM must try to address that and help client to change their opinion.

Thus, two text classification tasks should be performed in each turn of an MI dialogue, namely MITI classification and CT-ST classification. Some models are already introduced for these two, but their performance is not adequate for a robust

The Doctoral Centre for Safe, Ethical and Secure Computing (SEtS) at University of York

dialogue system [6], [7]. This can be attributed to the fact that they have not considered some sources of information that might help the classifier to make a better decision. For example, the history of dialogue, i.e the previous turns of a dialogue contain important clues which might increase or decrease the possibility of a specific label for the current turn. Another kind of information is the structural information inherent in the syntactic or semantic graph of sentences.

In this paper we will try to improve the performance of MITI classifiers with integrating the above information in the input of a state of the art text classifier. The history of dialogue and different type of graphs will be fed into the model in different ways to see which way yields a better result. To evaluate the effect, we perform our experiments on a relatively small but imbalanced dataset. We have the plan to examine the same idea on the task of CT-ST classification as well.

II. Related Works

For predicting MITI codes, different classification approaches such as SVM on linguistic patterns, topic modelling, and BERT language models are used [8], [9], and [10]. However, the classifier used in [6] can be considered as the state-of-the-art. It used a BERT-base architecture with weights initialized from a pre-trained LM, called RoBERTa, and trained three classifier instances based on three datasets. However, this model doesn't take into account the effect of previous turns in determining the label of current turn. Also, it only considers sentences' text as input whilst other representations of sentences such as graphs might contain some clues for distinguishing the label. For example dependency graph encodes the syntactic structure of sentence to some extent, and AMR graphs encode semantics, the relationships between words in a sentence according to their meanings. In this paper we are going to investigate the effect of integrating them into the above mentioned model.

III. Approach

In order to improve the performance of classifier, we intend to enrich the input of model in different ways. One is to inform the model about the text of previous sentences. This means that, in contrast to [6] in which only one sentence is given to model, we consider a history window size K, and provide also the last K sentences of the dialogue to the model. In such a setup, the model is able to condition the label, not only to what clinician is said in current utterance, but also to what has just passed in the previous utterances.

Another approach is to generate graph representation of input texts, and then feed the graph as input to the classification model. Since graphs are more structured than raw text, it is expected that the information inherent in their structure help the model to make better decisions. Different kinds of graph can be examined for this purpose. Dependency graph is one of them in which nodes correspond to tokens (words) of a text, and the edges represent the syntactic relationship between those words. An example of a dependency graph is shown in figure 1.

Fig. 1. An example of dependency graph for "You have the right to be sick and tired"

LOGIC format:

∃ w, b, g:
instance(w, want-01) ∧ instance(g, go-01) ∧
instance(b, boy) ∧ arg0(w, b) ∧
arg1(w, g) ∧ arg0(g, b)

AMR format (based on PENMAN):

```
(w / want-01
  :arg0 (b / boy)
  :arg1 (g / go-01
          :arg0 b))
```

GRAPH format:

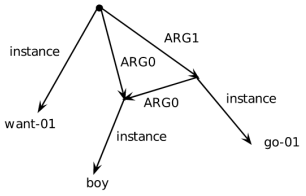

Fig. 2. An example showing how the sentence "The boy wants to go" is represented in AMR. Picture is taken from [12]

Another kind of graph is Abstract Meaning Representation (AMR) [11] in which leaf nodes are labelled with words in the sentence, non-leaf nodes represent higher-level concepts, and one of them is selected as the root that represents the focus of the sentence. The node labels come from either PropBank framesets, the English words, or some predefined keywords such as "world-region" and "monetary-quantity". Edges of the graph are also labelled with relations defined between words such as ARG0 and ARG1 that determine the subject and object of verbs, respectively. An example of a sentence represented with AMR is shown in Figure 2.

The above mentioned graphs can be fed into the model in different ways. One way to represent the graph in a textual encoding, and then pass that text as normal text to the model. Another way is to use Graph Neural Networks (GNN) and turn our text classification problem into a graph classification one, in which graph of a text is classified as one of 15 MITI labels. Such a GNN can independently act as our eventual classifier, or its output can be combined with the output of normal text classifier to form a two branch architecture, as depicted in Figure 3.

IV. Experimental Setup

Since we consider [6] as the baseline, we try to keep experimental setup as similar as possible to theirs. So we develop our model on top of code base published in [6]

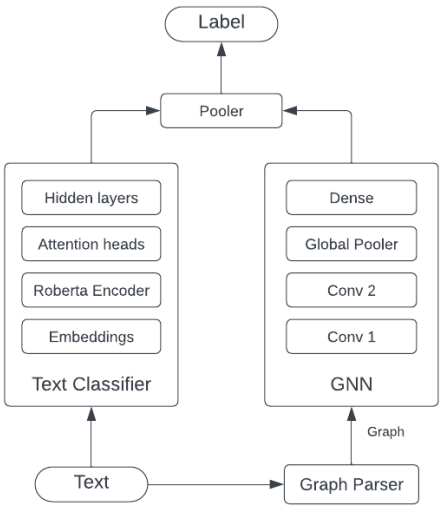

Fig. 3. The general flow of text classifier.

which is based on TensorFlow and a custom implementation of Transformers architecture [1].

We name the original code as "Boosting" and consider its performance as the base line. Then in order to verify effect of considering dialogue context, we run Boosting again but concatenate a window of size k=10 previous turns to the text of the current turn and then give it to the Boosting. We name this experiment as B(History). For the third and fourth experiments, we make a copy of Boosting model and feed the textual encoding of graphs into it, the outcome is then merged with output of Boosting using weighted average pooling to determine the final outcome. The overall arrangement is like Figure 3 but instead of both "Text classifier" and GNN, two copies of Boosting model is used. We call this architecture as B+(DPG) when dependency graphs are fed in, and B+(AMR) when AMR graphs are used. We also conduct two more experiments, almost similar to Figure 3, we use Boosting on the left hand side, and for the right hand side, using Spektral [13], we implement a simple Graph Convolutional Network containing two GCSConv's of size 256 and 512 respectively, followed by a GlobalAvgPool and a Dense layer. We name the overall model as B+GNN(DPG) and B+GNN(AMR) again depending on the type of graphs which are used as input.

Thus, overall five experiments are performed in addition to the original Boosting model. We kept all other configurations such as learning rate, optimizer, batch size, etc the same to maintain a more fair comparison.

A. Dataset

Three datasets are used in [6], out of which the first one (MI-Gold) is manually labeled and based on that, the other two are semi-automatically augmented. So we based our experiments only on MI-Gold. This dataset contains 2K dialogues, 17K

[1]https://github.com/anuradha1992/Boosting-with-MI-Strategy

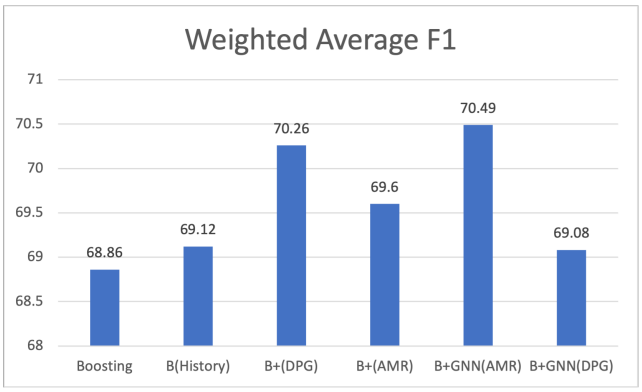

Fig. 4. The performance of different classification models.

utterances with imbalance ratio of 43:1, i.e. the number of most frequent label is 43 times the number of least frequent one.

Anno-MI is another publicly available dataset containing 133 MI sessions, more than 9.6k utterances, annotated with "Therapist Behaviour" (4 labels), and further with "Client Talk Type" (3 labels) which makes it suitable for CT-ST classification [14].

V. Evaluation

Since [6] reported its performance in terms of accuracy and weighted average F1 scores, we also used the same metrics for comparison. The results are shown in Figure 4. As it can be seen, all the experiments that combine Boosting with other models have superseded it, but the margin is not very significant. This can be due to the fact that our GNN architecture is not optimal. Up to now, we have only used a basic architecture and not so many other layouts, activation functions, etc are examined. Another reason might be the fact that the Boosting on the left hand side of the architecture is benefited from pre-trained weights that are already optimized on a larger corpus, whilst for the left counterpart, we are training it from scratch (initiated with random weights), so both sides of the model are not balanced in the extent of training.

VI. Conclusion

In this paper we argued about the importance of accurate speech act classification for the purpose of developing robust dialogue systems for mental healthcare. Particularly for MI, two classifiers are needed out of which one is targeted in this research (MITI codes classification) and the other is planned for future work (CT-ST classification). For the former, we considered [6] as state-of-the-art and developed two architectures on top of that. We then trained these two models on two kinds of graphs, dependency graphs and AMR. The results showed that integrating structural information of text into text classification has positive effect on performance but to a limited extent. For future works, with the aim of achieving a wider margin, we will try to train the GNN side of model

independently first, and once reached to a good accuracy, then pair it up with Boosting model. Since in the current experiments we didn't use the edge properties of the graphs, this is also planned to be done in near future.

REFERENCES

[1] W. R. Miller and S. Rollnick, *Motivational Interviewing: Helping People Change*. Guilford Press, Sep. 2012.

[2] R. M. Shingleton and T. P. Palfai, "Technology-delivered adaptations of motivational interviewing for health-related behaviors: A systematic review of the current research," *Patient education and counseling*, vol. 99, no. 1, pp. 17–35, 2016.

[3] D. Lenat and G. Marcus. (2023) Getting from Generative AI to Trustworthy AI: What LLMs might learn from Cyc. [Online]. Available: http://arxiv.org/abs/2308.04445

[4] T. B. Moyers, L. N. Rowell, J. K. Manuel, D. Ernst, and J. 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, Jun. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0740547216000143

[5] W. R. Miller, T. B. Moyers, Ernst, and P. Amrhein, "Motivational Interviewing Skill Code (MISC) 2.1," *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*, vol. 2, p. 50, Jan. 2008.

[6] A. Welivita and P. Pu, "Boosting Distress Support Dialogue Responses with Motivational Interviewing Strategy," May 2023. [Online]. Available: http://arxiv.org/abs/2305.10195

[7] Y. I. Nakano, E. Hirose, T. Sakato, S. Okada, and J.-C. Martin, "Detecting Change Talk in Motivational Interviewing using Verbal and Facial Information," in *Proceedings of the 2022 International Conference on Multimodal Interaction*, ser. ICMI '22. New York, NY, USA: Association for Computing Machinery, Nov. 2022, pp. 5–14. [Online]. Available: https://dl.acm.org/doi/10.1145/3536221.3556607

[8] V. Pérez-Rosas, R. Mihalcea, K. Resnicow, S. Singh, L. An, K. J. Goggin, and D. Catley, "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*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1128–1137. [Online]. Available: https://aclanthology.org/E17-1106

[9] X. Wei, "Automatic assessment of motivational interview with diabetes patients," Ph.D. dissertation, University of Birmingham, Aug. 2020. [Online]. Available: https://etheses.bham.ac.uk/id/eprint/12301/

[10] D. J. Min, V. Perez-Rosas, and R. Mihalcea, "Navigating Data Scarcity: Pretraining for Medical Utterance Classification," in *Proceedings of the 5th Clinical Natural Language Processing Workshop*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 59–68. [Online]. Available: https://aclanthology.org/2023.clinicalnlp-1.8

[11] K. Knight, B. Badarau, L. Baranescu, C. Bonial, M. Bardocz, K. Griffitt, U. Hermjakob, D. Marcu, M. Palmer, and T. O'Gorman, "Abstract meaning representation (amr) annotation release 3.0," 2021.

[12] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider, "Abstract Meaning Representation for Sembanking," in *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 178–186. [Online]. Available: https://aclanthology.org/W13-2322

[13] D. Grattarola and C. Alippi, "Graph Neural Networks in TensorFlow and Keras with Spektral [Application Notes]," vol. 16, no. 1, pp. 99–106, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9321429/

[14] Z. Wu, S. Balloccu, V. Kumar, R. Helaoui, E. Reiter, D. Reforgiato 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)*, May 2022, pp. 6177–6181.