# Exploiting Dialogue Structure for Task-oriented Psychiatric Interviews

Anonymous ACL submission

## Abstract

Task-oriented dialogues are used to extract information pieces needed to complete a particular operational goal. They can be divided into two main categories: slot filling tasks such as flight booking or product buying, and global evaluation tasks such as criminal interrogation or medical consultation. In this paper, we argue that discourse structure plays an important role in the latter case, and use automated depression level estimation as our experimental setup. In particular, we make the hypothesis that both patient and therapist dialogue acts individually contain relevant information towards the final evaluation task (depressed vs. non depressed), and define a novel multi-view architecture to combine both individual views. Experiments carried out on the DAIC-WOZ dataset show performance improvements over baselines and state-of-the-art models.

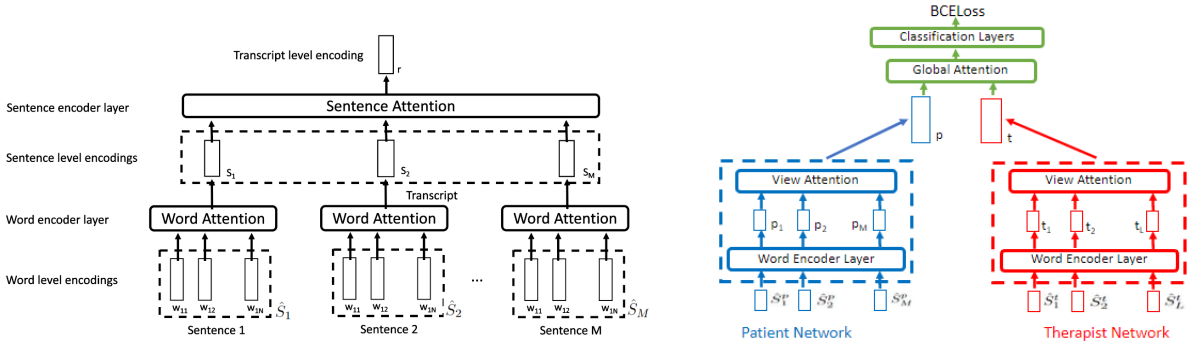
## 1 Introduction

Task-oriented dialogues are used to extract the information needed to achieve a particular conversational objective and can thus be divided into two main categories: form-filling tasks, and global assessment tasks. Within this context, most automated systems enable users to accomplish form-filling tasks, such as ticket booking, restaurant reservation, or customer support (Chen et al., 2017; Deriu et al., 2021). Other initiatives treat such dialogues at a global assessment level, especially in the healthcare domain, where systems should diagnose, monitor, assist, intervene or counsel patients (Valizadeh and Parde, 2022), as well as in the legal domain to a lesser extent (Hong et al., 2021).

An end-to-end dialogue system can be divided into three broad steps including natural language understanding, dialogue management and natural language generation (Chen and Gao, 2017). We focus our research on the natural language understanding part of the pipeline, structural language understanding in particular, leaving end-to-end system development as a long term objective. In particular, we focus on the global classification task of automated diagnosis (depressed or non-depressed) based on patient-therapist interview transcripts.

Depression is a serious mental disorder that affects millions worldwide, and an increasing curve is expected as a consequence of the current health crisis which has caused unemployment, stress and feeling of isolation among the population (Şimşir et al., 2022). Detection of depression is a challenging problem with patient-therapist interviews being the common practice to analyse a patient's mental health along with screening tools such as Personal Health Questionnaire (PHQ-8) depression scale. Throughout the literature, different strategies have been proposed for automated depression level estimations including multi-modal models (Ray et al., 2019; Qureshi et al., 2019), multi-task architectures (Qureshi et al., 2019, 2020), gender aware models (Qureshi et al., 2021; Bailey and Plumbley, 2021), hierarchical models (Mallol-Ragolta et al., 2019; Xezonaki et al., 2020), attention models with external knowledge (Xezonaki et al., 2020), feature-based solutions (Dai et al., 2021) and graph based solutions (Niu et al., 2021).

Despite this extensive list of research initiatives, ways to exploit the structure of an input transcript remain an unexplored research direction. In the context of structured interviews, questions asked by the interviewer are motivated towards an end task i.e., depression diagnosis, and therefore hold vital information. Most related works disregard the agent's identity, and by extension the identity of individual sentences as questions (therapist inputs) or answers (patient inputs). As such, these works force the model to understand the inter-dependencies within a sequence of unstructured utterances. In this paper, we argue that both patient and therapist dialogues hold relevant information, and propose multi-view architectures that exploit




(a) Hierarchical model

(b) Multi-view architectures

Figure 1: Overall models. (a) non-RNN based implementation of the hierarchical model; (b) Multi-view architectures where the view networks are outlined in red and blue and view fusion network is shown in green.

sentence identities to combine information from the two conversational agents for a better understanding of the input transcript. We conduct experiments on Distress Analysis Interview corpus - Wizard of Oz (DAIC-WOZ) dataset (Gratch et al., 2014) and show that multi-view architectures outperform both the baselines and the state-of-the-art models.

## 2 Related Work

Different architectures and strategies have been used throughout the literature to build models capable of estimating patients' depression level based on patient-therapist interviews. One promising research area is to combine inputs from different modalities into one learning model. Within this research direction, (Qureshi et al., 2019; Ray et al., 2019) explore the possibility of combining audio, visual and textual input features into a single architecture using attention fusion networks. Another interesting approach aims at combining different tasks that share some common traits thus following the multi-task paradigm. (Qureshi et al., 2020) propose to simultaneously learn both depression level estimation and emotion recognition on the basis that depression is a disorder of impaired emotion regulation, whereas (Qureshi et al., 2019) explore learning two tasks (regression and classification) simultaneously on the same dataset. Building on the success of hierarchical models for document classification, different studies (Mallol-Ragolta et al., 2019; Xezonaki et al., 2020) propose to encode patient-therapist interviews with hierarchical structures, showing boosts in performance. (Xezonaki et al., 2020) further extend their work and integrate affective information (emotion, sentiment, valence and psycho-linguistic annotations) from existing lexicons in the form of specific embeddings. Exploring a different research direction, (Qureshi et al., 2021) study the impact of gender on depression level estimation and build four different gender-aware models that show steady improvements over gender-agnostic models. Along the same line, (Bailey and Plumbley, 2021) study gender bias from audio features as compared to (Qureshi et al., 2021), who target textual information. Although most strategies rely on deep learning architectures, (Dai et al., 2021) propose building topic-wise feature vector based on a context-aware analysis over audio, video, text modalities.

## 3 Multi-view Strategy

In this paper, we propose a multi-view architecture that takes advantage of discourse structure (sentence identities in particular) to improve the learning capability of the model. We first present hierarchical models, used as base models, followed by a formal definition of the multi-view approach.

### 3.1 Hierarchical Model

Hierarchical models treat a patient-therapist interview as a hierarchy of intermediate representations (Mallol-Ragolta et al., 2019; Xezonaki et al., 2020), and are the current state-of-the-art for depression level estimation (binary classification). While these studies use recurrent neural networks (RNN), we define a non-RNN implementation of hierarchical models. This architectural choice is based on the findings of (Mohankumar et al., 2020), who show the limits of attention mechanisms over RNN encodings. Figure (1a) gives an overview of our hierarchical implementation, where *Word Attention* and *Sentence Attention* networks are defined as self-attention networks (Bahdanau et al., 2014). More details in (Xezonaki et al., 2020).

| Architectures | | macro F1 (Dev) | macro F1 Test | UAR (Dev) | UAR Test | Accuracy (Dev) | Accuracy Test | macro Precision (Dev) | macro Precision Test |
|---|---|---|---|---|---|---|---|---|---|
| Baselines | Patient | (0.6413) | 0.6429 | (0.6369) | 0.6361 | (0.6969) | **0.7608** | (0.6725) | 0.6584 |
| | Therapist | (0.8253) | 0.5818 | (0.8095) | 0.5803 | (0.8484) | 0.6521 | (0.8611) | 0.6184 |
| | Patient+Therapist | (0.7555) | 0.6053 | (0.7440) | 0.6004 | (0.7878) | 0.6739 | (0.7847) | 0.6250 |
| Multi-view | View-Global Attention | (0.6944) | 0.6811 | (0.6845) | 0.6674 | (0.7575) | 0.7391 | (0.7870) | **0.7252** |
| | Global Attention | (0.6857) | **0.7116** | (0.6785) | **0.7075** | (0.7272) | 0.7173 | (0.7083) | 0.6887 |
| | View Attention | (0.6944) | 0.6919 | (0.6845) | 0.6919 | (0.7575) | 0.6739 | (0.7870) | 0.6919 |

Table 1: Overall results over the DAIC-WOZ dataset. UAR stands for Unweighted Average Recall. We provide results for both development and test sets. The best model is chosen based on macro F1 over the development set.

### 3.2 Multi-view Architecture

We propose multi-view models that split the interview transcript into two views, patient view and therapist view, containing inputs from the respective conversational agents. Indeed, treating an interview as a sequence of sentences forces the model to understand the interactions between all pairs of sentences, which include a considerable portion of irrelevant interactions. Multi-view architectures utilize sentence identities to divide the input sequence into two views, thus avoiding irrelevant interactions like those between unrelated questions and answers. Our aim is to allow the model to focus only on inter-sentence interactions that exist within the two views in order to show the importance of sentence identities and how they can be used to filter input noise for improved performance.

Figure (1b) shows an overview of the proposed *View-Global Attention* multi-view architecture. In particular, the networks corresponding to the two views, i.e. *Patient Network* and *Therapist Network*, are instances of the hierarchical model defined in §3.1, and learn transcript level view representations $p$ (patient) and $t$ (therapist) of the corresponding input sequences. These representations are then combined through a *Global Attention* layer to encode weighted information from both *questions* and *answers* for the final task. *Sentence Attention* from the hierarchical model are renamed as *View Attention* in the multi-view architecture. Within this context we define three different configurations, which can be seen as an ablation study.

**View-Global Attention.** Within this configuration, both *View attention* layers and the *Global attention* layer are defined as self-attention networks.

**Global Attention.** In this configuration, *View Attention* layers are replaced by a simple averaging operation[1], while the *Global Attention* remains the same as in the View-Global Attention model.

**View Attention.** In this model, the *Global Attention* is replaced by a simple concatenation of the patient representation $p$ and the therapist representation $t$[2], while the *View Attention* layers remain the same as in the View-Global Attention model.

## 4 Dataset and Experimental Setups

**Dataset.** Automated depression detection is used as a specific use case to study multi-view architectures in the context of structured interview settings. In particular, we work on the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) dataset (Gratch et al., 2014) containing 189 clinical interviews designed to support diagnosis of the psychological distress conditions. The interviews are conducted by a virtual agent serving as the therapist, which is controlled by a human interviewer from another room (DeVault et al., 2014). The dataset contains inputs from different modalities (audio, visual and text) and has been annotated for a variety of verbal and non-verbal features with PHQ-8 acting as the final evaluation metric. Train, development and test sets contain 107, 35, 47 samples respectively, and have a class imbalance with 70% data points belonging to negative class (PHQ-8 score $\leq 10$) for binary classification.

**Implementation details.** In our experiments, we only target the text modality. For that purpose, we use pre-trained GloVe embeddings (300D) for word encodings (Pennington et al., 2014)[3]. Adam optimizer is utilized with a learning rate of $5 * 10^{-4}$ and the binary cross-entropy (BCELoss) is the final loss function. A dropout rate of 0.4 is applied. All implementations are done using PyTorch[4].

**Baseline models.** They are trained based on our implementation of hierarchical models (see Section §3.1) with three input configurations, *patient*, where only the patient speech acts are taken into account for the decision, *therapist*, where only the therapist questions are used for the decision pro-

[1]There is no attention information at sentence view level.

[2]There is no attention information at transcript level.

[3]https://github.com/stanfordnlp/GloVe

[4]Code will be released upon acceptance of the paper.

| Architectures | Modality | macro F1 | | UAR | |
|---|---|---|---|---|---|
| | | (Dev) | Test | (Dev) | Test |
| Raw Audio (2021) | A | (0.66) | - | - | - |
| SVM:m-M&S (2021) | T+V+A | (0.96) | 0.67 | - | - |
| HCAG (2021) | T+A | (0.92) | - | (0.92) | - |
| HCAN (2019) | T | (0.51) | 0.63 | (0.54) | 0.66 |
| HLGAN (2019) | T | (0.60) | 0.35 | (0.60) | 0.33 |
| HAN (2020) | T | (0.46) | 0.62 | (0.48) | 0.63 |
| HAN+L (2020) | T | (0.62) | 0.70 | (0.63) | 0.70 |
| HCAG+T (2021) | T | (0.77) | - | (0.82) | - |
| **Multi-view (Global)** | T | (0.68) | **0.71** | (0.68) | **0.71** |

Table 2: State-of-the-art results on DAIC-WOZ. T, V and A respectively stand for Text, Visual and Audio modalities.

cess, and *patient+therapist*, where the interview is taken as a sequence of questions (therapist inputs) and answers (patient inputs), similarly to (Xezonaki et al., 2020).

## 5 Analysis of the Results

Table 1 gives the detailed results for all configurations of baseline and multi-view architectures considered in this study. Figures show that multi-view architectures provide a strong way of combining inputs from patient-therapist interviews. In particular, multi-view models outperform the baselines in 3 out of 4 evaluation metrics, with the multi-view *Global Attention* model giving best overall results for 2 of the metrics, namely Macro F1 score and UAR. While this configuration provides best results in terms of precision/recall balance, the *View-Global Attention* model improves on precision but fails at recall on the test set.

Results obtained by taking only the questions as inputs (*Baseline Therapist*) verify the existence of information conveyed by the therapist, further strengthening the early conclusions made by (Xezonaki et al., 2020) over the General Psychotherapy Corpus[5]. Moreover, based on the results from Table 1 (*Baseline Patient+Therapist*) and the ones reported in (Xezonaki et al., 2020) (cf. §5 page 4559), we can argue that combining questions and answers as a sequence of sentences provides only marginal improvements at best when compared to using only patient inputs for the GPC corpus, and systematically downperforming in the case of the DAIC-WOZ dataset.

From the results obtained by the multi-view models compared to the ones with baseline hierarchical models, we can assess that multi-view architectures are a better alternative to process question-answer based interviews. Indeed, all multi-view architectures provide significant performance improvements over the *Baseline Patient+Therapist* model for 4 out of 4 evaluation metrics, highlighting the significance of retaining structural information of a dialogue in reducing the amount of noise and allowing the model to focus on relevant information during training.

Table 2 shows a comparison of our results with the state-of-the-art models. Our multi-view *Global Attention* model gives new state-of-the-art results in terms of macro F1 score and UAR, which are the standard evaluation metrics in the field, outperforming HAN+L (current SOTA) (Xezonaki et al., 2020), which includes a set of sentiment and emotion lexica as external knowledge. We can observe that interesting initiatives provide strong performance on the validation set for the text modality alone (HCAG+T (Niu et al., 2021)), but fail to generalize on the test set (results are omitted in their paper). Finally, it is known that multimodality plays an important role in the decision making process, as evidenced by performance differences between HCAG and HCAG+T models (Niu et al., 2021). As a consequence, we can hypothesize that there is a great margin for improvements of our multi-view model by integrating external knowledge and taking advantage of multimodality at least at patient level.

## 6 Conclusions and Future Works

In this paper, we define a novel multi-view architecture to process task-oriented dialogues for global evaluation, and show its effectiveness using depression classification as a use case. In particular, results obtained by the baseline models show the importance of therapists' questions and also the ineffectiveness of combining information from the two agents as a sequence of sentences. We further show that multi-view architectures steadily outperform comparable baselines, and state-of-the-art models relying on external knowledge (Xezonaki et al., 2020) or multiple modalities (Dai et al., 2021). Further improvements might be achieved by incorporating all the interactions that exist within an interview into the architecture, as well as including external knowledge as proposed by (Xezonaki et al., 2020). We also plan to extend this work and prove the results for similar task-oriented dialogue tasks such as candidate evaluation from job interviews, or criminal interrogation transcripts.

[5]Note that we were not able to purchase the GPC corpus from Alexander Street Press (many attempts) to replicate the study.

## References


Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate.

Andrew Bailey and Mark D. Plumbley. 2021. Gender bias in depression detection using audio features. In *29th European Signal Processing Conference (EUSIPCO)*, pages 596–600.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explorations Newsletter*, 19(2):25–35.

Yun-Nung Chen and Jianfeng Gao. 2017. Open-domain neural dialogue systems. In *Proceedings of the IJCNLP 2017, Tutorial Abstracts*. Asian Federation of Natural Language Processing.

Zhijun Dai, Heng Zhou, Qingfang Ba, Yang Zhou, Lifeng Wang, and Guochen Li. 2021. Improving depression prediction using a novel feature selection algorithm coupled with context-aware analysis. *Journal of Affective Disorders*, 295:1040–1048.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *International Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*, pages 1061–1068.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *9th International Conference on Language Resources and Evaluation (LREC)*, pages 3123–3128.

Jenny Hong, Catalin Voss, and Christopher Manning. 2021. Challenges for information extraction from dialogue in criminal law. In *1st Workshop on NLP for Positive Impact associated to the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 71–81.

Adria Mallol-Ragolta, Ziping Zhao, Lukas Stappen, Nicholas Cummins, and Björn W. Schuller. 2019. A hierarchical attention network-based approach for depression detection from transcribed clinical interviews. In *Interspeech (INTERSPEECH)*, pages 221–225. ISCA.

Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M. Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. 2020. Towards transparent and explainable attention models. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4206–4216.

Meng Niu, Kai Chen, Qingcai Chen, and Lufeng Yang. 2021. Hcag: A hierarchical context-aware graph attention model for depression detection. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4235–4239.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Syed Arbaaz Qureshi, Gaël Dias, Mohammed Hasanuzzaman, and Sriparna Saha. 2020. Improving depression level estimation by concurrently learning emotion intensity. *IEEE Computational Intelligence Magazine*, 15(3):47–59.

Syed Arbaaz Qureshi, Gaël Dias, Sriparna Saha, and Mohammed Hasanuzzaman. 2021. Gender-aware estimation of depression severity level in a multimodal setting. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Syed Arbaaz Qureshi, Sriparna Saha, Mohammed Hasanuzzaman, and Gaël Dias. 2019. Multitask representation learning for multimodal estimation of depression level. *IEEE Intelligent Systems*, 34(5):45–52.

Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg. 2019. Multi-level attention network using text, audio and video for depression prediction. In *9th International on Audio/Visual Emotion Challenge and Workshop (AVEC)*, page 81–88.

Mina Valizadeh and Natalie Parde. 2022. The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6638–6660.

Danai Xezonaki, Georgios Paraskevopoulos, Alexandros Potamianos, and Shrikanth Narayanan. 2020. Affective conditioning on hierarchical attention networks applied to depression detection from transcribed clinical interviews. In *Interspeech (INTERSPEECH)*, pages 4556–4560.

Zeynep Şimşir, Hayri Koç, Tolga Seki, and Mark D. Griffiths. 2022. The relationship between fear of covid-19 and mental health problems: A meta-analysis. *Death Studies*, 46(3):515–523.