Exploring the Influence of Dialog Input Format for Unsupervised Clinical Questionnaire Filling

Anonymous ACL submission

Abstract

In the medical field, we have seen the emergence of health-bots that interact with patients to gather data and track their state. One of the downstream application is automatic questionnaire filling, where the content of the dialog is used to automatically fill a pre-defined medical questionnaire. Answering questions from the dialog context can be cast as a Natural Language Inference (NLI) task and therefore benefit from current pre-trained NLI models. However, these models have not been generally trained on dialog input format, which may have an influence on their performance. In this paper, we study the influence of dialog input format on the task. Our results demonstrate that dialog preprocessing and content selection can significantly improve performance of zero-shot models.

1 Introduction

Recent pre-training and fine-tuning approaches have demonstrated strong performance gains on various natural language processing (NLP) tasks and benchmarks (Brown et al., 2020). However, fine-tuning still requires a considerable amount of task-specific data. Such data is not always available, and its collection can be very challenging.

One alternative is to use pre-trained models in a zero-shot setting. In particular, Toudeshki et al. (2021) showed that pretrained Natural Language Inference (NLI) models can be used to fill in questionnaires from dialogs in a zero-shot setting, i.e., without fine-tuning on task-specific data. They used dialogs without any pre-processing, however, while NLI models are typically trained on non-dialogic text.

In this paper, we propose different ways of transforming and selecting dialog content, and we analyse the impact of these operations on NLI-based questionnaire filling. Our experimental results demonstrate that, in a zero-shot setting, transforming and selecting dialog content yields significant improvements over a baseline which takes the full dialog content as input.

2 Related work

Given a text and a question, the purpose of a Machine Reading Comprehension (MRC) task is to derive the answer to that question from the input text (Zeng et al., 2020). Similarly, dialog-based, Automatic Questionnaire Filling can be viewed as an MRC task. Understanding multi-turn dialogs presents several challenges, however, as dialogs involve multiple speakers and intentions (Li et al., 2020), information may be imparted across multiple turns (Sun et al., 2019), topic shifts are frequent (Zhang et al., 2021) and utterances do not always appear in the form of complete sentences (Carbonell, 1983).

A simple approach for modeling a multi-turn dialog is to concatenate all turns (Zhang et al., 2019; Adiwardana et al., 2020). However, Zhang et al. (2018); Yuan et al. (2019) showed that turns-aware aggregation methods can achieve a better understanding of dialogs compared to considering all turns equally, in retrieval-based response selection for multi-turn conversations. Similarly, for multi-turn dialog MRC, turns-aware approach have been proposed which select turns in the conversation that are related to the input question. (Zhang et al., 2021) uses embedding-based similarity to select such turns while (Li et al., 2020) uses a pre-trained language model fine-tuned on NLI tasks. Their results showed that eliminating irrelevant turns effectively improves results.

Closest to our work, Toudeshki et al. (2021) showed that pre-trained NLI models can be used to fill in questionnaires from dialogs in a zero-shot setting. We depart from their work in that we propose different ways of transforming and selecting dialog content and investigate how this impact zero-shot, dialog-based, automatic questionnaire filling.

Dialog
bot: What is the most difficult for you about your sleep ? **patient:** I have back pain that prevents me from sleeping. **bot:** I'm sorry to hear that. How long have you had back pain? **patient:** Since I've been working out, I've had constant back pain at night. **bot:** Do you think pain can last for long? **patient:** I think it will stop once I stop playing sports. **bot:** Should we let time fix the pain? **patient:** My doctor thinks that I need to get used to doing sports and that the pain will disappear after a while.
Questionnaire **1.** My pain is a temporary problem in my life. CQ: (A) no **(B) yes** (C) NA ALS: (A) totally disagree (B) rather disagree (C) agree **(D) totally agree** (E) NA

Figure 1: An example of a dialog and a question from the PBPI Questionnaire, answered in CQ and ALS format

3 Automatic Questionnaire Filling (AQF)

Task. Given a dialog D and a questionnaire Q, the Automatic Questionnaire Filling task consists in providing an answer a_i for each question $q_i \in Q$.

Questionnaire. We consider two types of question: Close Questions (CQ) and Agreement Likert Scale (ALS) questions. Close questions have three possible answers (yes, no or Not Applicable, i.e. the dialog does not address the question) and ALS questions five (totally disagree, rather disagree, agree, totally agree, NA).

Data. We consider the questions listed in the Pain Beliefs and Perceptions Inventory (PBPI) questionnaire about pain beliefs and perception (Williams and Thorn, 1989). The questionnaire includes sixteen questions, all of them are formulated as declarative statements with multiple choice answers. PBPI questions along with ALS choices were used for ALS question type and respectively with CQ choices for CQ question type.

Dialogs. We evaluate our approach on ten dialogs collected by having the ComBot health bot (Liednikova et al., 2021) interact with a human agent. Dialog length varies from 13 to 55 turns and from 166 to 507 tokens, with 23.5 turns and 292.7 tokens in average.

4 Approach

We model question answering as an NLI task where the premise is derived from the dialog and the hypothesis from the question. We then derive answers from the NLI result.

Deriving an NLI hypothesis from a question and computing the answer. To derive an NLI hypothesis from a question, we represent questions as statements (E.g., "I have pain regularly" instead of "Do you have pain regularly?"). For Close Questions, the answer is then "yes" if NLI returns an entailment, "no" if it returns a contradiction and "NA" if it returns "neutral".

Similarly, for ALS, we represent each question as a statement and map the NLI result to agreement choices as follows. If "neutral" has the highest score, the answer is "NA". Else, the contradiction score is subtracted from the entailment score. The subtraction result lies in a range of (-1,1) which is uniformly divided into 5 segments corresponding to the 5 ALS answer types.

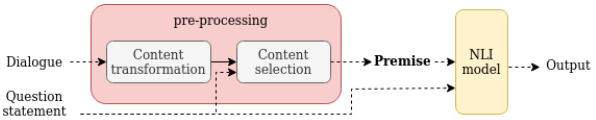

Figure 2: Dialog pre-processing schema

Deriving an NLI Premise from the dialog. The NLI premise is derived from the input dialog using first, Content Transformation and second, Content Selection. We describe these operations in the next section.

Model Architecture. Given a question and a dialog, our model, illustrated in Figure 2, answers the question in three steps as follows. First, the input dialog is transformed and dialog content is selected. This creates a premise for NLI. Second, NLI is applied to determine the entailment relation between this premise and the question (converted to a statement). Finally, the NLI results are used to compute the answer as described above. For NLI, we use Roberta large (Liu et al., 2019) [1] fine-tuned on the MNLI dataset (Williams et al., 2018).

5 NLI-oriented Dialog Pre-processing

We consider different options to transform and select dialog content. We then explore how each combination of options impacts AQF performance.

[1] https://huggingface.co/roberta-large-mnli

5.1 Content transformation

Null Transformation (CT_{null}). A (transformation) baseline where we simply concatenate the turns of the input dialog.

Summary (CT_{sum}) . Pairs of non overlapping, adjacent turns are summarized, and the resulting summaries are concatenated. In this way, the input dialog is transformed into a sequence of two-turn summaries. We also tried summarizing the whole dialog in one go but found that applying summarization on each two turns rather than on the whole dialog gives better results.We use the **BART-large** model[2] (Lewis et al., 2020) fine-tuned on the News summarization corpus XSUM (Narayan et al., 2018) and on the dialog summarization corpus SAMSum (Gliwa et al., 2019).

Long Answers (CT_{answer}). In information seeking dialog, adjacent turns often are question-answer pairs. Based on this observation, we map each pair of non overlapping, adjacent turns in the dialog into a single declarative sentence assuming that the first turn is a question (e.g., "Which drug did you take?"), the second is a short answer to that question (e.g., "Doliprane") and the sentence derived from the mapping is a long answer to the question (e.g., I took Doliprane). To learn this mapping, we fine-tune T5 (Raffel et al., 2019), a pre-trained encoder-decoder model, on two datasets of (question, incomplete answer, full answer) triples, one for wh- and one for yes-no (YN) questions. For wh-questions, we use 3,000 entries of the dataset consisting of (question, answer, declarative answer sentence) triples gathered by (Demszky et al., 2018) using Amazon Mechanical Turk workers. For YN questions, we used SAMSum corpus (Gliwa et al., 2019) which contains short dialogs in chit-chat format. We created 1,000 (question, answer, full answer) triples by automatically extracting (YN question, answer) pairs from this corpus and manually associating them with the corresponding declarative answer.

This fine-tuned model was applied to each two subsequent turns of the input dialogs, and the resulting declarative sentences were then concatenated to form the declarative transform of the whole dialog.

[2] https://huggingface.co/Salesforce/bart-large-xsum-samsum

5.2 Content selection

The transformation operations described in the previous section yield sequences of dialog turns, two-turn summaries or full answers. We call these "input units" and consider three ways of pre-selecting the input units that will be used as premise when testing for entailment.

Null Content Selection (CS_{null}) A (content selection) baseline where the premise is the concatenation of all the input units produced by the content transformation operations (dialog turns, sequence of two turn summaries, sequence of full form answers).

Unit-Based (CS_{units}). Each question is assessed against each input item. Given an input sequence I_n of length n, the answer a_i to a question q is then determined by aggregating the resulting entailment probabilities as follows:

- $a_i = NA$ if for all input items $i \in I_n$, the NA probability is highest.
- $a_i = Yes$ (resp. $a_i = No$) if for at least one item $i \in I_n$, the Yes (resp. No) probability is highest and the highest Yes (resp. No) probability is higher than the highest No (resp. Yes) probability.

Similarity (CS_{sim}). For each question q, we select a subset of input units that are semantically similar to q. We encode question and input units using SBERT paraphrase-distilroberta-base-v2[3] (Reimers and Gurevych, 2019) and compute *cosine similarity* for each (q, input unit) pair. We then select items whose similarity score is higher than 0.5, concatenate them and use the result as the NLI premise.

NLI (CS_{nli}). For each question q in the questionnaire, we select the input units that are related to q using the NLI model (the same model that has been used for the task). Specifically, we select sentences which have entailment or contradiction score higher than 0.5. All selected sentences are then concatenated to form the NLI premise.

6 Results

We evaluate our approach using macro and weighted F1 score. Figures 3 and 4 show the results

[3] https://huggingface.co/sentence-transformers/paraphrase-distilroberta-base-v2

	CQ					ALS						
	NA	YES	NO	macro	weighted	NA	TD	RD	A	TA	macro	weighted
***CT**$_{null}$*												
CS$_{null}$	0.25	0.07	0.31	0.21	0.17	0.21	0.17	0.19	0.05	0.04	0.13	0.11
CS$_{units}$	0.42	0.70	0.59	0.57	0.63	**0.42**	0.48	0.0	0.0	0.72	0.32	0.40
CS$_{sim}$	0.35	0.69	0.24	0.43	0.50	0.41	0.16	0.15	0.16	0.69	0.31	0.38
CS$_{nli}$	0.38	0.74	0.67	0.60	0.67	**0.42**	0.53	0.12	0.05	**0.73**	**0.37**	**0.44**
***CT**$_{sum}$*												
CS$_{null}$	0.27	0.15	0.30	0.24	0.21	0.17	0.27	**0.20**	0.09	0.13	0.17	0.16
CS$_{units}$	0.35	0.43	0.57	0.45	0.47	0.35	0.37	0.07	0.11	0.44	0.27	0.30
CS$_{sim}$	0.30	0.42	0.04	0.25	0.29	0.29	0.06	0.09	0.12	0.49	0.21	0.26
CS$_{nli}$	0.32	0.48	0.61	0.47	0.49	0.34	0.41	0.06	0.19	0.54	0.31	0.36
***CT**$_{answer}$*												
CS$_{null}$	0.30	0.32	0.31	0.31	0.31	0.29	0.32	0.17	0.09	0.22	0.22	0.22
CS$_{units}$	0.42	0.71	0.63	0.59	0.64	**0.42**	0.48	0.0	0.06	0.70	0.33	0.41
CS$_{sim}$	0.36	0.73	0.35	0.48	0.56	0.39	0.30	0.08	**0.29**	0.70	0.35	0.42
CS$_{nli}$	**0.45**	**0.79**	**0.67**	**0.64**	**0.70**	**0.42**	**0.61**	0.0	0.05	**0.73**	0.36	**0.44**

Table 1: F1-Scores for closed (CQ) and agreement Likert scale (ALS) question type; TD - totally disagree, RD - rather disagree, A - agree, TA - totally agree. CT: content transformation, CS: content selection.

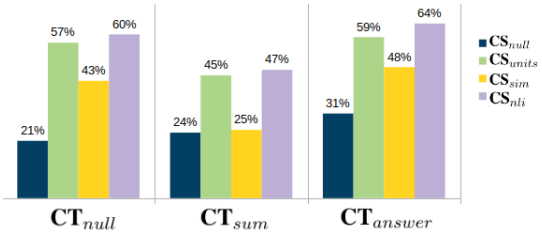

Figure 3: F1 macro average for Close Questions

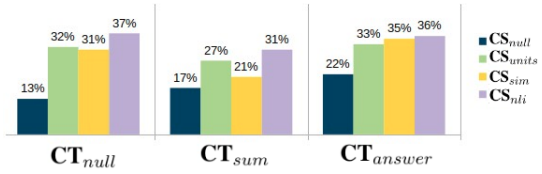

Figure 4: F1 macro average for ALS questions

in graphical form. Table 1 shows the breakdown of the results by answer type.

On both types of questions, content selection yields significant improvement over the null content selection baseline (doubling the F1 score for some configurations) while the long answer transform, which merges pairs of adjacent dialog turns into statements (CT$_{answer}$), consistently yields the best results. For instance, for ALS questions, we see that using the CT$_{answer}$ transformation together with NLI-based content selection (CS$_{nli}$) multiplies the macro F1 score by three and the weighted F1 score by four compared to the CT$_{null}$, CS$_{null}$ baseline.

From the three content transformation methods, summarization has the lowest performance which is likely due to loss of important information and hallucinations.

The difference between the null (CS$_{null}$) and the unit-based (CS$_{units}$) content selection approaches suggests that the model performs more accurately when the premise is shorter. In both approaches, the entire content is considered for final decision. However, for CS$_{null}$, the model receives all content at once, while for CS$_{units}$, the model handles the input content, one item at a time.

Finally, we see that agreement answers (Yes, Totally agree) have the highest accuracy (over 70%) in both question types which suggests that the NLI model is better at confirming rather than rejecting a statement.

7 Conclusion

In this paper, we studied how dialog pre-processing can impact the task of filling medical questionnaires based on patient-bot interactions. Our experimental results show that converting pairs of adjacent turns to sentences and selecting input units based on their entailment relation with the question can significantly enhance performance, thereby reducing the need for model adaptation using few-shot learning.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang,

Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Jaime G Carbonell. 1983. Discourse pragmatics and ellipsis resolution in task-oriented natural language interfaces. In *21st Annual Meeting of the Association for Computational Linguistics*, pages 164–168.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Junlong Li, Zhuosheng Zhang, and Hai Zhao. 2020. Knowledgeable dialogue reading comprehension on key turns. *arXiv preprint arXiv:2004.13988*.

Anna Liednikova, Philippe Jolivet, Alexandre Durand-Salmon, and Claire Gardent. 2021. Gathering information and engaging the user combot: A task-based, serendipitous dialog model for patient-doctor interactions. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 21–29.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Farnaz Ghassemi Toudeshki, Philippe Jolivet, Alexandre Durand-Salmon, and Anna Liednikova. 2021. Zero-shot clinical questionnaire filling from human-machine interactions. In *MRQA*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

David A. Williams and Beverly E. Thorn. 1989. An empirical assessment of pain beliefs. *Pain*, 36(3):351–358.

Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2019. Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 111–120.

Changchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. 2020. A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. *Applied Sciences*, 10(21):7640.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. *arXiv preprint arXiv:1806.09102*.

Zhuosheng Zhang, Junlong Li, and Hai Zhao. 2021. Multi-turn dialogue reading comprehension with pivot turns and knowledge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1161–1173.

A Experiment time estimation

The experiments were conducted with a laptop having Intel® Core™ i7-10610U CPU @ 1.80GHz * 8 and NVIDIA Quadro P520.

B Questionnaire

In this section, we provide the PBPI questionnaire statements.

nb.	Question
1	No one is able to tell me why it hurts.
2	I thought my pain could be healed, but now I'm not so sure.
3	There are times when it doesn't hurt.
4	My pain is difficult for me to understand.
5	My pain will always be there.
6	I am in constant pain.
7	If it hurts, it's only my fault.
8	I don't have enough information about my pain.
9	My pain is a temporary problem in my life.
10	I feel like I wake up with pain and fall asleep with it.
11	I am the cause of my pain.
12	There is a way to heal my pain.
13	I blame myself when it hurts.
14	I can't understand why it hurts.
15	One day, again, I won't have any pain at all.
16	My pain varies in intensity but it is always present with me.

Table 2: List of questions in PBPI questionnaire