

Zero-Shot Clinical Questionnaire Filling From Human-Machine Interactions

Farnaz Ghassemi Toudeshki^{&,#}, Philippe Jolivet[&],
Alexandre Durand-Salmon[&], Anna Liednikova^{&†}
& ALIAE

[#] IDMC, Université de Lorraine

[†] Université de Lorraine, LORIA

{farnaz.ghassemi, philippe.jolivet,
alexandre.durand-salmon, anna.liednikova}@aliae.io

Abstract

In clinical studies, chatbots mimicking doctor-patient interactions are used for collecting information about the patient’s health state. Later, this information needs to be processed and structured for the doctor. One way to organize it is by automatically filling the questionnaires from the human-bot conversation. It would help the doctor to spot the possible issues. Since there is no such dataset available for this task and its collection is costly and sensitive, we explore the capacities of state-of-the-art zero-shot models for question answering, textual inference, and text classification. We provide with a detailed analysis of the results and propose further directions for clinical questionnaire filling.

1 Introduction

Chatbots in healthcare can be used to collect information about the user with different purposes: treatment adherence, monitoring, patient support program, patient education, behavior change, diagnosis (Car et al., 2020).

Considering monitoring, patient-doctor conversations have mainly been used for the automation of medical records’ creation through the extraction of clinical entities such as symptoms, medications, and their properties (Du et al., 2019), generating reports (Finley et al., 2018) and summaries (Zhang et al., 2018). Surprisingly, the task of filling clinical questionnaires received less attention.

Patients fill standard questionnaires during each medical visit, which frequency is usually several weeks. Performing this task in an automated way based on serendipitous talk (through a chatbot) opens the opportunity to get updated information more regularly and then monitor the patient more closely and in a seamless way.

Ren et al. (2020a) were the first to introduce a questionnaire filling task as a classification problem, with the targets in the form of symptom

phrases. Though, there is plenty of questionnaires that consist of full meaningful assertions or questions. The difference between questionnaire filling and slot filling is in the complexity of the questions that require machine reading comprehension (MRC) and question answering (QA).

The goal of a typical MRC task is to process a (set of) text passage(s) and then to answer questions about the passage(s). Though usually, multi-choice answer options are semantically different, in the case of questionnaires, the answers are often on the same scale (agree-disagree, often-rare).

More concretely, we make the following contributions:

- a clinical questionnaires’ categorization based on questions and answers types
- data collection schema for filling questionnaires for 5 question types: open questions (OQ), closed questions (CQ), agreement Likert-scale (ALS), frequency Likert-scale (FLS) and visual analogue scale (VAS)
- analysis of question answering (QA), natural language inference (NLI), and zero-shot text classification (ZeroShot-TC) state-of-the-art models performance for the mentioned questions types

2 Related work

Four formats are commonly used for posing questions and answering them: Extractive (EX), Abstractive (AB), Multiple-Choice (MC), and Yes/No (YN). UnifiedQA (Khashabi et al., 2020) is a single pre-trained QA model, that performs well across 20 QA datasets spanning 4 diverse formats.

Demszky et al. (2018) proposed a sentence transformation model which converts question-answer pairs into their declarative forms that allows to solve the task with NLI models. On the other hand, (Yin et al., 2019) shows how ZeroShot-TC models

may be successfully used as NLI models, being partially inspired by the way (Obamuyide and Vlachos, 2018) transformed relation classification task into NLI problem.

Questions in multiple-choice RACE (Lai et al., 2017) require holistic reasoning about the text, involving people’s mental states and attitudes and the way they express them. Mishra et al. (2020) described reasoning categories present in RACE and showed that if given passage is a dialogue, NLI model performs better than QA model.

Ren et al. (2020b) presented a new medical Information Extraction task which uses questionnaires to convert medical text data into structured data. Their work is based on neural network classifier which makes selection among given options for closed-end type questions to fill out one complete questionnaire using only one model.

To our best knowledge, it is a first work to address the problem of filling in clinical questionnaires based on dialogue history. In Section 3 we are going to discuss the most common question types in questionnaires, including Likert scale which wasn’t addressed before in clinical natural language processing. Also in Section 6 we show state-of-the-art models’ ability to solve this task.

The task of filling questionnaires from user-bot dialogue history is a very specific sub-field of MRC in NLP and as a result, reduces the chance of availability of such data for training/fine-tuning current models or later for evaluation, specially in medical domain. Both dialogues and answered questionnaires based on them are required that makes it difficult to collect such data on a large scale. This low resource setting is conducive to zero-shot approaches.

3 Question types

Khashabi et al. (2020) categorized existing QA datasets into four categories based on the format of answers: Extractive (EX), Abstractive (AB), Multiple-Choice (MC), and Yes/No (YN). In the following section we would like to introduce also Likert scale and Visual analogue scales (VAS) type of answers which is very common for clinical questionnaires. Sinha et al. (2017) came to the conclusion in their review that Visual analogue scale (VAS) and numerical rating scale (NRS) were the best adapted pain scales for pain measurement in endometriosis. Also, VAS is often used in epidemiologic and clinical research to measure the intensity

Question type	Questionnaire
Open Question (OQ)	Morin (Morin, 1993) QCD (Daut, 1983) MOS-SS (Sherbourne and Stewart, 1991)
Closed Question (CQ)	QPC (C. Thomas-Antérion, 2004) EPICES (Bihan et al., 2005)
Agreement Likert-scale (ALS)	TSK (Kori, 1990) PBPI (Williams and Thorn, 1989) TAS-20 (Bagby et al., 1994) LOT-R (Scheier et al., 1994) IEQ-CF (Sullivan, 2008) JCK (KARASEK, 1985)
Frequency Likert-scale (FLS)	MOS-SS (Sherbourne and Stewart, 1991) MBI (Maslach et al., 1997) PCL-S (Weathers et al., 1993) HAD (Zigmond and Snaith, 1983) SF-12 (Ware Jr et al., 1996)
Visual Analogue Scale (VAS)	QCD (Daut, 1983) Dallas (Lawlis et al., 1989) LEEDS (Parrott and Hindmarch, 2004) FABQ-W (Waddell et al., 1993)

Table 1: Clinical questionnaires and their types of questions

or frequency of various symptoms (Paul-Dauphin, 1999).

For each question type present in medical questionnaire, we provide examples of questionnaires and question in Table 1.

4 Task and Data

4.1 Task

Given a human-healthbot dialog history D and a set of questions q_i extracted from a questionnaire Q , the task is to determine the correct answer a_i to each question q_i , including the ’not mentioned’ option, i.e. the possibility that the dialog does not address that question.

4.2 Chatbot

To create a chatbot for interactions, we followed the ComBot ensemble (Liednikova et al., 2021). The conversation always starts with the opening ’What is the most difficult for you about your sleep?’. We made sure that there is no intersection between questions of the bot and questionnaire ques-

tions, due to the point of the research for answering questions implicitly addressed in the dialog. Since creating a dialog following the questionnaire topics is a very resource consuming task, we decided to mitigate the risk of the system proposing undesired direction of the conversation with a rejection function. If a bot reply isn't consistent with the dialog history, the user can reject it and receive the next best candidate to continue the conversation.

4.3 Questionnaires

For experiments, we have chosen three questionnaires that are semantically close to the topics of the chatbot model: Morin (OQ), PBPI (ALS, VAS, CQ), Mos-ss (FLS). For PBPI questionnaire we ask annotators to give the answers in three format types at the same time: CQ, ALS, VAS. We provide statistics in Table 2. The full list of the questions could be found in Appendix, the answer options could be found in Tables 5 and 6.

4.4 Data collection

We asked 10 annotators to interact with the chatbot one time for each of the three questionnaires. So in total, we collected 30 dialogues and their corresponding question answers.

The annotators were first asked to read the questionnaire so that they could guide the interaction with the health bot maximizing the number of questions addressed during the dialog. Then the annotators were asked to fill in the questionnaire based on their conversation and to select 'Not mentioned' (NA) option if the current question couldn't be answered from the dialog history. For PBPI questionnaire we ask annotators to give the answers in three format types at the same time: CQ, ALS, VAS. You can find screenshot of interfaces in Appendix and answer options in Tables 5 and 6.

To ensure the reliability of collected data, we conducted a double annotation with adjudication. We ask two people to fill in the questionnaires based on the collected dialogues. So, totally for each question of each dialog we have three answers (one from the author of the dialog and two from other annotators). Table 4 demonstrates final agreement between two annotators engaged in double annotation, as well as between two annotators and initial participants. In case of disagreement between annotators, the third person (adjudicator) decides the final label. These ground truth labels are used for evaluation of the models later.

During annotation, we came with some definition of classes that helped the annotators to come to agreement. We consider that the user would totally agree with the questionnaire statement if this statement or its paraphrase is explicitly mentioned in text, otherwise if there are phrases that fully or partially support this statement we annotate it with agree label. The same rules are applied to give disagreement and total disagreement label for contradictory statements.

Due to the high level of complexity of the task, we enrolled high-educated volunteers from our professional network. We made sure that participants were well aware of the context of the work and they were all properly compensated. On average, performing a complete conversation with chatbot took 20 minutes from participants, and about 15 minutes for answering a questionnaire from a dialog history.

Since the data and questionnaires were initially in French, we have translated them with DeepL¹ to English and to run the models.

Q Type	Questionnaire	Nb. of Q	Nb. of A
OQ	Morin	22	inf
CQ	PBPI	16	3
ALS	PBPI	16	5
FLS	Mos-ss	10	7
VAS	PBPI	16	11

Table 2: Questionnaires statistics: number of questions and number of answer options

	MOS-SS	Morin	PBPI
# turn	24.4	34.5	23.5
# tokens	259.1	403.2	269.2
# nb of tokens / turn	11.2	11.9	12.2
# Q answered	76%	59.3%	85.6%
# uniq tokens	140.8	201	147.6

Table 3: Statistics of dialogues for each questionnaire

5 Models

In this section, we present the models that can be used in zero-shot setting with selected question types.

QA model UnifiedQA (Khashabi et al., 2020)² is a single pre-trained QA model, that performs

¹<https://www.deepl.com/fr/translator>

²<https://github.com/allenai/unifiedqa>

OQ	CQ	ALS	FLS	VAS
two annotators				
0.81 (BertScore)				
0.72 (Rouge-1)	0.81	0.70	0.67	0.31
two annotators + user				
0.75 (BertScore)				
0.67 (Rouge-1)	0.67	0.47	0.56	0.20

Table 4: Inter-annotator agreement, using Kappa score for closed-ended questions and F1 score for Open question type

question type	choice	ratio
CQ	yes	55%
	no	30.6%
	NA	14.4%
ALS	totally agree	35%
	agree	19.4%
	rather disagree	11.8%
	totally disagree	19.4%
	NA	14.4%
FLS	all the time	5.6%
	most of the time	11.9%
	a good part of the time	5%
	sometimes	12.5%
	rarely	7.5%
	never	5%
	NA	15%
VAS	10	34.4%
	9	1.2%
	8	3.1%
	7	11.9%
	6	3.8%
	5	0%
	4	0.6%
	3	3.1%
	2	8.1%
	1	1.2%
	0	18.1%
	NA	14.4%

Table 5: Statistics of number of different choices for each question type

well across 20 QA datasets spanning 4 diverse formats. Fine-tuning this pre-trained QA model into specialized models results in a new state of the art on 10 factoid and commonsense QA datasets, establishing UnifiedQA as a strong starting point for building QA systems. In our experiments, we used UnifiedQA-t5-3b version.

Applied to the following question types:

- **OQ** The dialog history and question are provided in format of NarrativeQA dataset (Kočíský et al., 2017)
- **CQ and ALS** To fit format of MCTest dataset (Richardson et al., 2013), we transformed questionnaire statements to question form, by

changing first pronoun to second and adding "Do you agree with that" at the beginning. Otherwise, the model tends to choose NA option.

- **FLS** The dialog history and question are provided in format of MC Test dataset (Richardson et al., 2013)

NLI model We use DeBERTa V2 xlarge model (He et al., 2020)³ fine-tuned with MNLI dataset (Williams et al., 2018) for NLI task. We pass to the model concatenated dialog history as premise and question in declarative form as hypothesis. The output is probabilities for three classes: Entailment, Contradiction, Neutral.

Applied to the following question types:

- **CQ** A question is transformed into a statement and treated as hypothesis. Entering the premise and hypothesis to the model, we choose the class with the highest score as the final answer. Entailment class is considered as 'yes', contradiction as 'no' and neutral as 'NA'.
- **ALS** Premise is a dialogue history, hypothesis is a questionnaire statement. If probability for neutral class was higher than contradiction and entailment class, we consider that the dialogue doesn't contain relevant information to the question and give NA label. Otherwise, we take probability of Contradiction with negative sign and sum up it with probability of Entailment with positive sign. The resulting score lies in the interval of (-1,1). We uniformly divide this interval into N segments, where N is a number of options on Linkert-scale (usually 4 or 5).
- **FLS** For each question, we enter the model with dialogue as the premise and concatenation of one of the frequency scales with statement format of question (hypothesis = freq + question_statement_format) as hypothesis. We treat transformed question-answer combinations as distinct hypotheses, as presented in (Trivedi et al., 2019). Among frequency choices, we choose the one which has the highest entailment score. If none of them has the entailment score higher than 50%, we consider that the dialogue doesn't contain relevant

³<https://github.com/microsoft/DeBERTa>

information to the question and give NA label. Otherwise, we select the frequency scale with the highest entailment score.

- **VAS** If score for neutral is highest, then the selected output would be NA. Otherwise, we subtract the entailment score from contradiction. The result would be in range (-1,1). To map the result in range (0,10), we add value 1 to the subtraction result and then multiply it with 5 (shown in equation 1).

$$value = (ent - cont + 1) * 5 \quad (1)$$

where *ent* and *cont* are the predicted probabilities of entailment and contradiction classes.

ZeroShot-TC model We use Bart-large model (Lewis et al., 2019)⁴ for zero-shot text classification trained on MNLI corpus (Williams et al., 2018). In this setting, we pass concatenated dialog history as a context and formulate target labels as filled templates, so that the input data could be closer to entailment format. Then we add one more target label 'NA' to consider the situation when the answer is not mentioned in the dialog. The model provides probability scores for each candidate, and the candidate with the highest probability would be chosen as the final answer.

Applied to the following question types:

- **CQ** We transform a question into the statement beginning with "I agree that" for "yes" choice or with "I disagree that" for "no" choice.
- **ALS** The candidates are formed following the template "I *agreement_option* that *questionnaire statement*", where agreement options are totally disagree, disagree, agree, totally agree.
- **FLS** We transform a question into the statement and add in the beginning the frequency option to form the candidates.
- **VAS** Candidates are defined as the same as ones in CQ. Model outputs probability scores for each class. By having these scores for each label and map them to entailment/contradiction/neutral labels, we continue transforming the results to VAS scale by doing the same process explained previously for VAS using NLI model.

⁴<https://huggingface.co/facebook/bart-large-mnli>

6 Evaluation

Since for each question type we had different output, we used different evaluation scores.

For evaluating the performance of UnifiedQA on open-question types, we have used two common evaluation metrics: ROUGE (Lin, 2004) and BERTscore Zhang et al. (2019). ROUGE counts how many n-grams in the generated response matches the n-grams in the reference answer. Since we used abstractive mode of UnifiedQA, generated outputs might have different n-grams from reference ones. Therefore, using BERTscore might show a more realistic evaluation perspective since it computes the semantic similarity between generated and referenced answers based on embeddings.

We evaluate closed questions (CQ), agreement Likert scale (ALS), frequency Likert scale (FLS) and visual analogue scale (VAS) with macro and weighted F1 score.

The results are shown in the Table 7 for open-ended question type and in Table 8 for closed-ended ones. Because UnifiedQA generates answer even for unmentioned questions, we report the results for both all questions (mentioned and not mentioned questions) and for just mentioned ones.

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

7 Results and discussion

Detecting unanswerable questions Considering all open questions, we can see from Table 7 that the scores are considerably low, 0.38 for ROUGE and 0.55 for BERT. On the other hand, if we calculate these scores only for open questions being mentioned in the dialog (according to the user answers), the scores are almost 2 times better. Results indicate the high performance of UnifiedQA model for answering mentioned questions and its lack to distinguish given the context if the question is answerable or not.

Impact of number of choices Table 8 shows the performance of SOTA NLI and ZeroShot-TC models for answering closed-ended question types (CQ, ALS, FLS, VAS). Comparing results for different question types can tell us that number of multiple-choices in each question type has a great impact on final results. Closed question type with 3 choices has the highest results and on the other hand, VAS with 11 choices has the lowest performance.

Q type	Answers	NLI		ZeroShot-TC	
CQ	yes, no, NA	Premise	dialogue history	Context	dialogue history
		Hypothesis	There are times when I don't have pain.	Targets	<i>I agree that</i> there are times when I don't have pain, <i>I disagree that</i> there are times when I don't have pain., NA
		Model output	entailment(0.5), contradiction(0.3), neutral(0.2)	Model output	I agree that there are times when I don't have pain.
		Final output	yes	Final output	yes
ALS	totally disagree, rather disagree, agree, totally agree, NA	Premise	dialogue history	Context	dialogue history
		Hypothesis	There are times when I don't have pain.	Targets	<i>I totally disagree that</i> there are times when I don't have pain., <i>I rather disagree that</i> there are times when I don't have pain., NA
		Model output	entailment(0.5), contradiction(0.3), neutral(0.2)	Model output	I totally disagree that there are times when I don't have pain.
		Final output	agree	Final output	totally disagree
FLS	all the time, most of the time, a good part of the time, sometimes, rarely, never, NA	Premise	dialogue history	Context	dialog history
		Hypothesis	{freq_scale} I got the amount of sleep I needed.	Targets	<i>all the time</i> I got the amount of sleep I needed., <i>most of the time</i> I got the amount of sleep I needed., ..., NA
		Model output	entailment score for each freq. scale	Model output	rarely I got the amount of sleep I needed.
		Final output	choosing freq. scale which has the highest entailment score	Final output	rarely
VAS	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, NA	Premise	dialogue history	Context	dialog history
		Hypothesis	I got the amount of sleep I needed.	Targets	<i>I agree that</i> there are times when I don't have pain, <i>I disagree that</i> there are times when I don't have pain., NA
		Model output	entailment(0.5), contradiction(0.3), neutral(0.2)	Model output	I agree that there are times when I don't have pain.(0.5), I disagree that there are times when I don't have pain.(0.3), NA(0.2)
		Final output	6	Final output	6

Table 6: Examples of input and output for textual inference with Deberta and zero-shot classification with Bart-large for closed, agreement, frequency and VAS scale questions

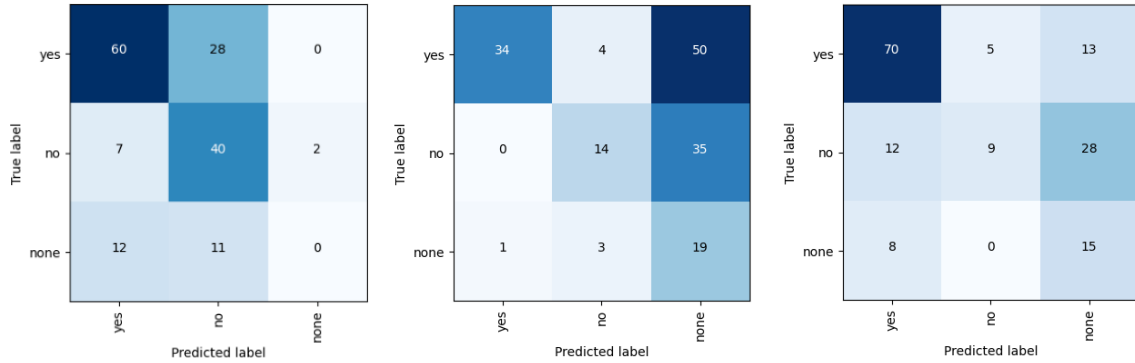


Figure 1: Confusion matrix for CQ using QA (left), NLI (center) and ZeroShot-TC (right) models

Metric	All	Answered
ROUGE	0.38	0.63
BERT	0.55	0.93

Table 7: Scores for zero-shot evaluation of OQ type

The table 8 also indicates the superiority of ZeroShot-TC for CQ and FLS questions types than NLI. After comparing confusion matrices provided for the NLI and ZeroShot-TC models, we can ob-

serve that the NLI model has a high tendency to give NA (neutral) class as output, while this is not the case for the ZeroShot-TC model. On the other hand, the inability of ZeroShot-TC model to correctly predict extreme choices (totally agree/totally disagree) for ALS question type (figure 2) has led to the lower performance of this model in comparison with NLI.


Model \ Question type	metric	CQ	ALS	FLS	VAS
Random (Baseline)		0.33	0.25	0.14	0.09
UnifiedQA-t5-3b	macro F1	0.44	0.13	0.29	
	weighted F1	0.58	0.12	0.32	
deberta-v2-xlarge-mnli	macro F1	0.417	0.240	0.158	0.064
	weighted F1	0.470	0.262	0.192	0.104
facebook/bart-large-mnli	macro F1	0.484	0.166	0.220	0.04
	weighted F1	0.575	0.136	0.262	0.03

Table 8: Scores for zero-shot evaluation for question types: CQ - closed question, ALS - agreement Likert-scale, FLS - frequency Likert-scale, VAS - Visual Analogue Scale

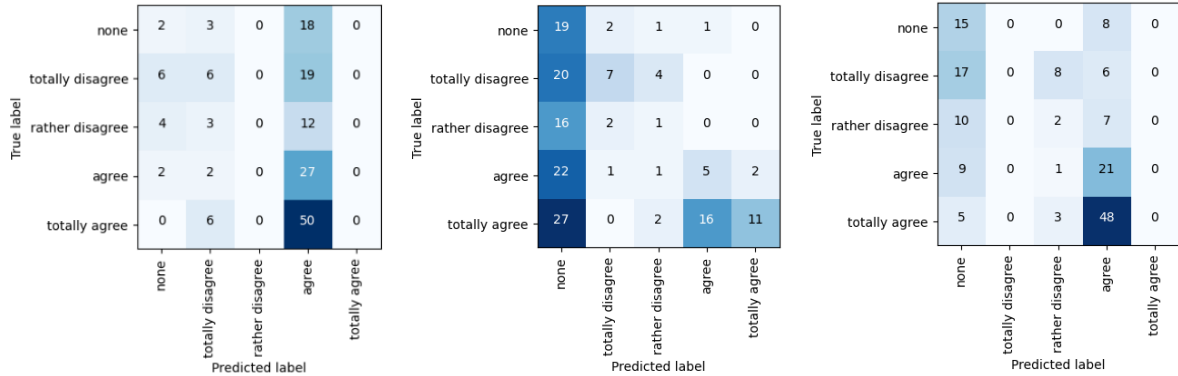


Figure 2: Confusion matrix for ALS using QA (left), NLI (center) and ZeroShot-TC (right) models

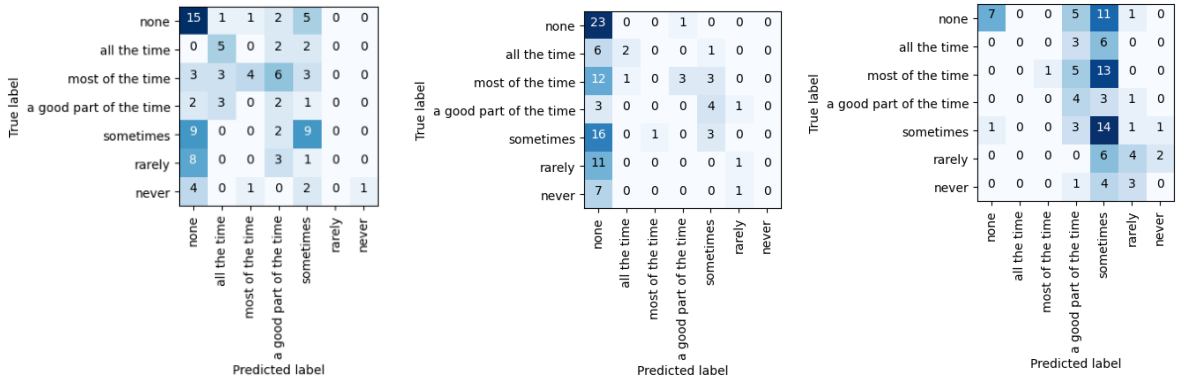


Figure 3: Confusion matrix for FLS using QA (left), NLI (center) and ZeroShot-TC (right) models

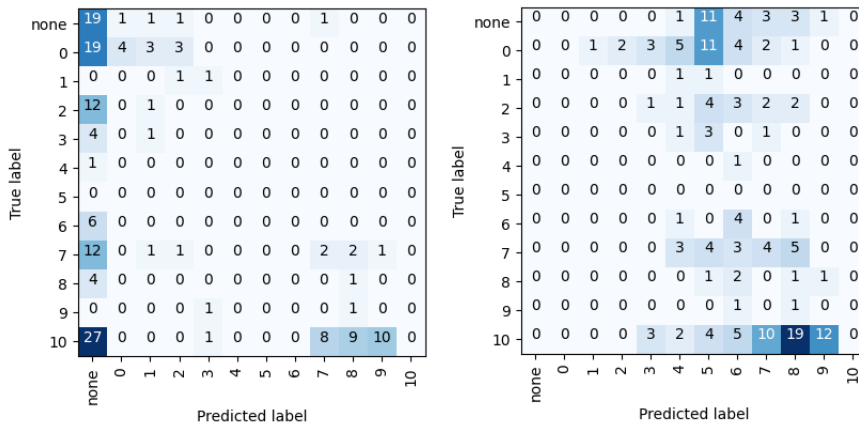


Figure 4: Confusion matrix for VAS using NLI (left) and ZeroShot-TC (right) model

Measuring agreement is the most challenging task From Table 8 and Figures 2 and 4 we can see that predicting answers for agreement scale is the most challenging task. Since the answer option isn't semantically different enough to facilitate models choice, the probabilities for target classes don't help with selecting the correct level of agreement. In future work, we would like to explore other approaches, such as multi-hop reasoning and argument mining.

Importance of text input From our experiments, we can derive that the models are sensitive to the input text format and noise. Also, the different models are sensitive to the different data preprocessing technics. They may include using speaker tags, punctuation cleaning, selecting a subset of input text on some criteria. Such experiments should show what kind of preprocessing may boost performance without unnecessary data collection and training. These results may be a contribution to green and sustainable NLP.

8 Conclusion

In this paper, we introduced the approaches for clinical questionnaire filling in a zero-shot setting for the five question types most used in clinical studies. Also, we describe the data collection process for their evaluation. With the results, we show that this task is not easy to solve.

There is a type of questions which are more well-known in multi-hop reading comprehension. For such type of questions, there is a need to properly integrate multiple pieces of evidence to answer them. Song et al. (2018) investigates graph convolutional network (GCN) and graph recurrent network to perform evidence integration.

As future work, we plan to take advantage of graph convolutional networks (GCNs) to improve the textual entailment for questionnaire filling. It would be possible due to enriching the model with knowledge graphs both in the open domain and close domain (medical).

Another direction to explore is text transformations. For example, we plan to transform statement-question into cloze form by masking the most important word and providing it as one option of the answers, use common-sense knowledge and graph structure for better reasoning.

Ethical considerations

Regarding Regulation (EU) 2017/745, described software is intended for general uses, even when used in a healthcare environment, it is intended for uses relating to lifestyle or well-being that do not constitute any a medical prediction and medical prognosis function without doctors validation or correction.

Acknowledgments

The authors would like to warmly thank members of the Aliae company for their support and helpful feedback during the project. They would also like to thank Claire Gardent for valuable discussions at earlier stages of this work and constructive feedback.

References

- R.michael Bagby, James D.a. Parker, and Graeme J. Taylor. 1994. [The twenty-item toronto alexithymia scale—i. item selection and cross-validation of the factor structure.](#) *Journal of Psychosomatic Research*, 38(1):23–32.
- Hélène Bihan, Silvana Laurent, Catherine Sass, Gérard Nguyen, Caroline Huot, Jean Jacques Moulin, René Guegen, Philippe Le Toumelin, Hervé Le Clésiau, Emilio La Rosa, Gérard Reach, and Régis Cohen. 2005. [Association among individual deprivation, glycemic control, and diabetes complications.](#) *Diabetes Care*, 28(11):2680–2685.
- S. Honoré-Masson G. Berne P.H. Ruel B. Laurent C. Thomas-Antérion, C. Ribas. 2004. [Le questionnaire de plainte cognitive \(qpc\) : outil de dépistage de la plainte des sujets présentant une maladie d'alzheimer ou un mci.](#) *Revue Neurologique*, 1027(4502):5–75.
- Lorraine Tudor Car, Dhakshenya Ardhithy Dhinakaran, Bhone Myint Kyaw, Tobias Kowatsch, Shafiq Joty, Yin Leng Theng, and Rifat Atun. 2020. [Conversational agents in health care: Scoping review and conceptual analysis.](#)
- Cleeland C. S. Flanery R. C. Daut, R. L. 1983. [Development of the wisconsin brief pain questionnaire to assess pain in cancer and other diseases.](#) *Pain*, (17):197–210.
- Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. [Transforming question answering datasets into natural language inference datasets.](#)
- Nan Du, Kai Chen, Anjuli Kannan, Linh Tran, Yuhui Chen, and Izhak Shafran. 2019. [Extracting symptoms and their status from clinical conversations.](#) In

- Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 915–925, Florence, Italy. Association for Computational Linguistics.
- Greg P. Finley, Erik Edwards, Amanda Robinson, Najmeh Sadoughi, James Fone, Mark Miller, David Suendermann-Oeft, Michael Brenndoerfer, and Nico Axtmann. 2018. An automated assistant for medical scribes. In *INTERSPEECH*.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. [Deberta: Decoding-enhanced bert with disentangled attention](#).
- R KARASEK. 1985. [Job content questionnaire user’s guide](#). *Department of Work Environemnt*.
- Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. [Unifiedqa: Crossing format boundaries with a single qa system](#).
- Miller R.P. Todd D.D. Kori, S.H. 1990. [Kinesiophobia: A new view of chronic pain behavior](#). *Pain Management*, (3):35–43.
- Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2017. [The narrativeqa reading comprehension challenge](#).
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. [RACE: Large-scale ReAding comprehension dataset from examinations](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.
- G FRANK Lawlis, RAMON Cuencas, DAVID Selby, and CE McCoy. 1989. The development of the dallas pain questionnaire. an assessment of the impact of spinal pain on behavior. *Spine*, 14(5):511–516.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- 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, Online. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Christina Maslach, Susan E Jackson, and Michael P Leiter. 1997. *Maslach burnout inventory*. Scarecrow Education.
- Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Li, Pavan Kapanipathi, and Kartik Talamadupula. 2020. [Reading Comprehension as Natural Language Inference: A Semantic Analysis](#).
- Charles M Morin. 1993. *Insomnia: Psychological assessment and management*. Guilford press.
- Abiola Obamuyide and Andreas Vlachos. 2018. [Zero-shot relation classification as textual entailment](#). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium. Association for Computational Linguistics.
- A. Parrott and I. Hindmarch. 2004. The leeds sleep evaluation questionnaire in psychopharmacological investigations—a review. *Psychopharmacology*, 71:173–179.
- Guillemin F. Virion J. M. Briançon S. Paul-Dauphin, A. 1999. [Bias and precision in visual analogue scales: a randomized controlled trial](#). *American journal of epidemiology*, 150(10):1117–1127.
- Jiangtao Ren, Naiyin Liu, and Xiaojing Wu. 2020a. [Clinical questionnaire filling based on question answering framework](#). *International Journal of Medical Informatics*, 141:104225.
- Jiangtao Ren, Naiyin Liu, and Xiaojing Wu. 2020b. [Clinical questionnaire filling based on question answering framework](#). *International Journal of Medical Informatics*, 141:104225.
- Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. [MCTest: A challenge dataset for the open-domain machine comprehension of text](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.
- Michael F. Scheier, Charles S. Carver, and Michael W. Bridges. 1994. [Distinguishing optimism from neuroticism \(and trait anxiety, self-mastery, and self-esteem\): A reevaluation of the life orientation test](#). *Journal of Personality and Social Psychology*, 67(6):1063–1078.
- Cathy Donald Sherbourne and Anita L. Stewart. 1991. [The mos social support survey](#). *Social Science Medicine*, 32(6):705–714.
- Sarthak Sinha, Amanda J. Schreiner, Jeff Biernaskie, Duncan Nickerson, and Vincent A. Gabriel. 2017. [Treating pain on skin graft donor sites: Review and clinical recommendations](#). *Journal of Trauma and Acute Care Surgery*, 83(5):954–964.
- Linfeng Song, Zhiguo Wang, Mo Yu, Yue Zhang, Radu Florian, and Daniel Gildea. 2018. Exploring

graph-structured passage representation for multi-hop reading comprehension with graph neural networks. *arXiv preprint arXiv:1809.02040*.

Adams H. Horan S. Mahar D. Boland D. Gross R. Sullivan, M.J.L. 2008. The role of perceived injustice in the experience of chronic pain and disability: Scale development and validation. *Journal of Occupational Rehabilitation*, (18):249–61.

Harsh Trivedi, Heeyoung Kwon, Tushar Khot, Ashish Sabharwal, and Niranjan Balasubramanian. 2019. Repurposing entailment for multi-hop question answering tasks. *arXiv preprint arXiv:1904.09380*.

Gordon Waddell, Mary Newton, Iain Henderson, Douglas Somerville, and Chris J Main. 1993. A fear-avoidance beliefs questionnaire (fabq) and the role of fear-avoidance beliefs in chronic low back pain and disability. *Pain*, 52(2):157–168.

John E Ware Jr, Mark Kosinski, and Susan D Keller. 1996. A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. *Medical care*, pages 220–233.

Frank W Weathers, Brett T Litz, Debra S Herman, Jennifer A Huska, Terence M Keane, et al. 1993. The ptsd checklist (pcl): Reliability, validity, and diagnostic utility. In *annual convention of the international society for traumatic stress studies, San Antonio, TX*, volume 462. San Antonio, TX,;

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.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. [Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach](#). 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 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. [Bertscore: Evaluating text generation with bert](#). *arXiv preprint arXiv:1904.09675*.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. [Learning to summarize radiology findings](#). In *EMNLP 2018 Workshop on Health Text Mining and Information Analysis*.

Anthony S Zigmond and R Philip Snaith. 1983. The hospital anxiety and depression scale. *Acta psychiatrica scandinavica*, 67(6):361–370.

A Questionnaires

nb.	Question
1	Did you have the impression that your sleep was not calm (moving constantly, feeling tense, talking, etc., while you were sleeping)?
2	Did you get enough sleep to feel rested when you woke up in the morning?
3	Did you wake up short of breath or with a headache?
4	Have you felt a cloudy or drowsy mind during the day?
5	Have you had difficulty falling asleep?
6	Have you woken up in your sleep and had difficulty falling back to sleep?
7	Did you have trouble staying awake during the day?
8	Have you snored in your sleep?
9	Did you take naps (5 minutes or more) during the day?
10	Did you get the amount of sleep you needed?

Table 9: List of questions in MOS-SS questionnaire

B User interface for annotation

[Having calm sleep]

1. Did you have the impression that your sleep was not calm (moving constantly, feeling tense, talking, etc., while you were sleeping)?

None
 All the time
 Most of the time
 A good part of the time
 Sometimes
 Rarely
 Never

[Enough sleep to feel rested]

2. Did you get enough sleep to feel rested when you woke up in the morning?

None
 All the time
 Most of the time
 A good part of the time
 Sometimes
 Rarely
 Never

[Wake up with headache or out of breath]

3. Did you wake up short of breath or with a headache?

None
 All the time
 Most of the time
 A good part of the time
 Sometimes
 Rarely
 Never

[Being drowsy during the day]

4. Have you felt your mind foggy or drowsy during the day?

None
 All the time
 Most of the time
 A good part of the time
 Sometimes
 Rarely

Figure 5: MOS-SS questionnaire

nb.	Question
1	What time do you usually go to bed on weekday evenings?
2	What time is your final wake-up call in the morning?
3	What time do you usually get up during the week?
4	What time do you go to bed on your days off?
5	What time do you get up on your days off?
6	How soon do you fall asleep after turning off the light?
7	How many times a night do you wake up on average?
8	How long do you spend waking up between your first sleep and your final awakening?
9	How many hours per night do you sleep on average?
10	Why do you wake up at night? (pain, noise, child, nightmare, spontaneous awakening, others)
11	What medicine are you taking or were you taking and at what dose?
12	How many nights per week do you currently take the medicine?
13	When did you start taking the medicine?
14	When was the last time you took the medicine?
15	How long have you suffered from insomnia?
16	When was the first time you had trouble sleeping?
17	Did your insomnia start gradually or suddenly?
18	Were there any stressful events that could be linked to the onset of your insomnia? (death, divorce, retirement, family or professional problem, ..)
19	How many times a week do you exercise?
20	How much coffee, tea or Coca-Cola do you consume per day?
21	How many cigarettes do you smoke per day?
22	How many glasses of beer, wine or alcohol do you drink per day?

Table 10: List of questions in Morin questionnaire

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 11: List of questions in PBPI questionnaire

The image shows a digital form for the Morin questionnaire. It contains several sections with questions and corresponding input fields:

- [Time of going to bed]**: Question 1.5: "What time do you usually go to bed on weekday evenings?"
- [Time of final awakening in the morning]**: Question 1.6: "What time is your final wake-up call in the morning?"
- [Time of getting up]**: Question 1.7: "What time do you usually get up during the week?"
- [Time of going to bed on off days]**: Question 1.8: "What time do you go to bed on your days off?"
- [Time of getting to bed on off days]**: Question 1.9: "What time do you get up on your days off?"
- [Time takes to fall asleep]**: Question 1.12: "How soon do you fall asleep after turning off the light?"
- [Average number of waking up while sleeping]**: Question 1.13: "How many times a night do you wake up on average?"
- [Total awakenings times at night]**: Question 1.14: "How long do you spend waking up between your first sleep and your final awakening?"

Figure 6: Morin questionnaire

[Misunderstanding about my pain by others]

1- No one is able to tell me why it hurts.

None Totally disagree Rather disagree Agree Totally agree

None No Yes

None 0 1 2 3 4 5 6 7 8 9 10

[Healing uncertainty]

2- I thought my pain could be healed, but now I'm not so sure.

None Totally disagree Rather disagree Agree Totally agree

None No Yes

None 0 1 2 3 4 5 6 7 8 9 10

[Not feel hurt]

3- There are times when I don't have pain.

None Totally disagree Rather disagree Agree Totally agree

None No Yes

None 0 1 2 3 4 5 6 7 8 9 10

[Understanding of my pain by myself]

4- My pain is difficult for me to understand.

None Totally disagree Rather disagree Agree Totally agree

None No Yes

None 0 1 2 3 4 5 6 7 8 9 10

[Losing hope]

5- My pain will always be there.

None Totally disagree Rather disagree Agree Totally agree

None No Yes

None 0 1 2 3 4 5 6 7 8 9 10

Figure 7: PBPI questionnaire