Clinical Questionnaire Filling from Human-machine Interactions

Presenter: Farnaz Ghassemi Toudeshki
Supervisors: Anna Liednikova, Philippe Jolivet

November 22, 2021

Table of contents

Introduction

Question Types in Clinical Questionnaires

Evaluation Dataset

Study Zero-shot Models

Exploring the Influence of Input Format

Proposing a Graph-based NLI Model

Introduction

About Aliae

- Aliae is a French company based in Nancy that develops a new type of therapeutic support with the goal of improving patients' quality of life (QoL) between 2 medical visits.
- Aliae uses cutting edge tools and methods to collect, and analyze patient's insights through a chatbot.
- Aliae's chatbot, called ComBot [1], is designed to interact with patient in his daily living to understand how the patient feels and translates it into actionable data (physicians' monitoring & reporting, therapeutic education).

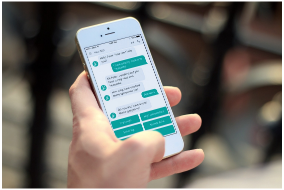
Introduction

- Aliae's main focus is chronic pain, which can impact negatively QoL (quality of sleep, level of anxiety, the mood, social interactions, level of activity).
- Sleep disorder is the topic we focus on for this research, as we have access to physicians' expertise.

Introduction

The Task

- Closed questionnaires is the traditional way to collect QoL data.
- Filling in these questionnaires takes time, is very repetitive and may require medical knowledge to understand it.



- Extracting key points from patient-bot day-to-day chats can be quite beneficial to get a more accurate view of patient's state **evolution**.

Introduction (cont.)

- The **Goal** of this research is to study the task of "Automatically filling medical questionnaires from patient-bot interactions".
- The task differs from multiple-choice QA, because set of choices in medical questionnaires can be very semantically close to each other.

Dialogue
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. (A) totally disagree (B) rather disagree (C) agree **(D) totally agree** (E) NA **2.** No one is able to tell me why it hurts. **(A) totally disagree** (B) rather disagree (C) agree (D) totally agree (E) NA 3. ...

Introduction

Studies

1. Firstly, we investigate the **capabilities of state-of-the-art zero-shot models** for the task. This is due to the lack of relevant datasets and also the difficulties of data collection.
2. Secondly, we **explore the influence of dialogue input format** and experiment several dialogue pre-processing approaches and show their impact on final results.
3. And finally, we **propose a graph-based NLI model** for the task.

Question Types in Clinical Questionnaires

Question Types in Clinical Questionnaires

Common question types in medical questionnaires:

- **Open question**: answer in text format
- **Closed question**: answer is either yes or no
- **Likert scale question**: answer ranges from one extreme attitude to another
- **Visual Analogue scale**: answer is placed on a continuum of values

Question Types in Clinical Questionnaires (cont.)

Question type	Choices
Open questions (OQ)	-
Closed questions (CQ)	Yes, No, NA
Agreement likert-scale (ALS)	Totally disagree, Rather disagree, Agree, Totally agree, NA
Frequency Likert-scale (FLS)	All the time, Most of the time, A good part of the time, Sometimes, Rarely, Never, NA
Visual analogue scale (VAS)	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, NA

Table 1: Question types

In this study, we focus on following question types.

Evaluation Dataset

Evaluation Dataset

There is a **lack of data to test methods** for answering medical questionnaires from dialogue. This urged us to **take the initiative of collecting such data**.

Chatbot

- To create a chatbot for interactions, we followed the ComBot ensemble [1].
- Combot is a health-bot designed to interact with people who suffer from insomnia and sleep disorder, as well as to track the user's status on a regular basis.

Evaluation Dataset

Questionnaires

We have chosen three questionnaires that are semantically close to the topics of the chatbot model: Morin, PBPI, Mos-ss.

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

Table 2: Questionnaires used for data collection

Evaluation Dataset

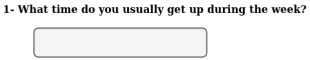


Figure 1: A sample question of Morin questionnaire

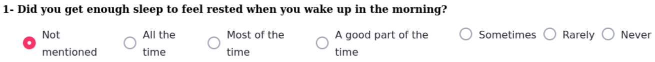


Figure 2: A sample question of Mos-ss questionnaire

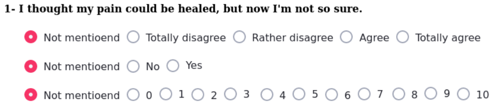


Figure 3: A sample question of PBPI questionnaire

Evaluation Dataset

Data Collection

- For each of the three questionnaires, we asked 10 participants to engage with the chatbot once.
- After the conversation, the participants were then asked to fill in the questionnaire based on the information presented during the chat.
- To ensure the reliability of collected data, we conducted a double annotation with adjudication.
- The ground truth labels will be used to evaluate the models.

1st Study:
Study Zero-shot Models

Study Zero-shot Models

The following NLP methods were chosen to investigate the capabilities of SOTA zero-shot models for the task:

1. **QA**
   - The task of providing an answer in response to the question.
   - Model used: **UnifiedQA-t5-3b** [2]
2. **NLI**
   - The task of determining whether or not one statement can be deduced from another.
   - Model used: **deberta-v2-xlarge-mnli** [3]
3. **ZeroShot-TC**
   - The task of classifying a text between any provided labels.
   - Model used: **bart-large-mnli** [4]

Results

Metric	All	Answered
ROUGE	0.38	0.63
BERT	0.55	0.93

Table 3: Scores for zero-shot evaluation of OQ type

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 4: Scores for zero-shot evaluation of CQ, ALS, FLS, VAS

Study Zero-shot Models (cont.)

- Good performance of UnifiedQA in answering mentioned questions.
- The UnifiedQA's inability to differentiate between mentioned and unmentioned questions.
- Number of multiple-choices in each question type has a great impact on final results.
- Predicting level of agreement is the most challenging task.
- NLI model has a high tendency to give NA (neutral) as output.
- Models are sensitive to the text input format.

2nd Study: Exploring the Influence of Input Format

Exploring the Influence of Input Format

- The aim of this section is to explore the impact of dialogue input format on zero-shot models.
- For this study, we concentrate on zero-shot NLI model which is more related to our task.
- We show how different pre-processing approaches lead to different results.

Exploring the Influence of Input Format

Dialogue pre-processing includes two steps:

1. Content transformation
   - For transforming dialogue to declarative form
2. Content selection
   - For selecting premise out of main content

(various approaches were investigated for each step)

Exploring the Influence of Dialog Input Format

Results

- Entering NLI model with declarative content, instead of dialogue, considerably improves the performance.
- NLI model can discriminate better between different classes when premise is shorter.
- The NLI model performs better at confirming rather than rejecting a statement.

3rd Study: Proposing a Graph-based NLI Model

Graph-based NLI Model

- We **propose a graph-based NLI approach** for the task of automatically filling questionnaires from dialog histories.
- By converting text inputs to graphs and using a graph-based model, we can enrich inputs by **domain knowledge** as well as various types of **linkages between sentence units**.
- The approach contains 2 main steps: (1) **Graph construction** out of premise and hypothesis, and (2) **Graph classification model** using R-GCN framework.

Graph-based NLI Model

Graph construction

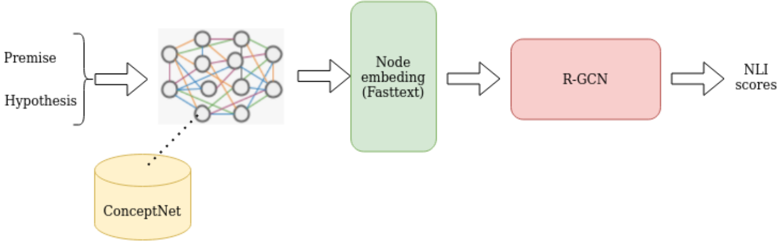
- Forming graph out of premise and hypothesis
- Using **dependency parsing** in graph structure (to focus on one-to-one correspondences between single words)
- Enriching the graph with **ConceptNet** (to better characterize the concepts in the input texts)

Graph Representation Learning

- Using **Relational Graph Convolutional Network (R-GCN)**[5] for graph encoding
- Using **Fasttext** for initial node embeddings

Graph-based NLI Model

Model structure



Graph-based NLI Model

Training Dataset: SNLI corpus [6]

- 23,192 train (4% of whole snli train-set)
- 10,000 dev.
- 10,000 test

Results on SNLI

train set	dev. set	test set
79.7	73.8	72.1

Table 5: Accuracy of model on SNLI

Graph-based NLI Model

Future work

- Adapt the model for dialogue format premise
- Using domain-specific knowledge base / ontology
- See the effectiveness of other initial node embedding approaches
- Investigate other graph representations/ structures

Thank you!

Bibliography

[1] A. Liednikova, P. Jolivet, A. Durand-Salmon, and C. Gardent, "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*, 2021, pp. 21–29.

[2] D. Khashabi, S. Min, T. Khot, *et al.*, "Unifiedqa: Crossing format boundaries with a single qa system," *arXiv preprint arXiv:2005.00700*, 2020.

[3] P. He, X. Liu, J. Gao, and W. Chen, *Deberta: Decoding-enhanced bert with disentangled attention*, 2020. arXiv: `2006.03654 [cs.CL]`.

Bibliography (cont.)

[4] M. Lewis, Y. Liu, N. Goyal, *et al.*, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[5] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *European semantic web conference*, Springer, 2018, pp. 593–607.

[6] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *arXiv preprint arXiv:1508.05326*, 2015.