Multi-party Conversation to Query-Response Pair Conversion Annotation Guidelines

I. Introduction

The purpose of this annotation is to convert multi-party online conversations into a set of multiple query-response pairs. Thereby the resulting dataset may be used for response generation for multimodal queries.

As an example, the original data may look appear as follows:

Table 1. Example of original online multiparty conversation. Besides the original author, other authors may provide questions or comments that respond to the original content or to subsequent turns.

turn	author	content
1	U001	刚看的病人,病史20年 
2	U002	寻常狼疮
4	U003	原帖由 U002 于 2028-8-29 15:49 发表 寻常狼疮 检查真菌与结核.

5	U004	支持寻常狼疮
6	U005	牛皮癣?

The constructed final data has the following format:

Table 2. Example of converted query-response pairs. A single query with one or more images is specified. Responses should only refer to the query and not refer to other turn content.

type	author	content
query	U001	刚看的病人,病史20年 
response	U002	寻常狼疮
	U003	寻常狼疮 检查真菌与结核.
	U004	寻常狼疮
	U005	牛皮癣?

To achieve this, the data is processed with both human annotation and an automatic algorithm.
The human annotation includes:

- ***Section II: Conversation Exclusion*** - include/exclude entire conversations
- ***Section III: Turn Labeling and Exclusion*** - include/exclude individual turns
- ***Section IV: Turn Content-Editing*** - edit turn content to make each reply context-free (e.g. not rely on information from other author replies)

Described in **Section V: Automatic Query-Response Construction**, an automatic algorithm then removes empty text content turns, removes non-advice replies, and merges consecutive turns by the same authors, to construct the final form of the dataset.

II. Conversation Exclusion

Table 3. Exclusion cases by content type

content_type	Description of applicable exclusion criteria per content
Original Turn	Exclude if: - The text is not related to a clinical problem[1] - The original turn content already has the answer[2] - The original turn content has no text content[3]
Replies	Exclude if: - No replies from other authors that give either request for more information or professional advice
Image	Exclude if: - Image includes genitalia - Image includes face/identifiable features - Image has annotations (e.g. user-draw arrows or circles)

Note: If any of these cases are met, we do not need to do the annotations below.

Examples:

[1] Exclude the conversations where doctors are sharing an interesting case they have seen. In this case, replies will not be informative.

username	content	label
w****y	今天头一次见到痛风石，如果不对大家指正。大家有治疗方法吗？	EXCLUDE_THREAD
q****d	学习了，谢谢	
针****原	患者得有望闻问切等各种病史症状才能下诊断，只是这样看不能确诊吧	
福*堂	这个很典型，但 更需要实验室检查，更严肃。	
w****s	痛风的确诊要根据临床症状、体征，同时结合实验室检查等。严谨！	
z***m	本人也是痛风，不过没那么严重，哎，自己不给自己治病	
f****9	很典型的图片，学习了，谢谢！	

[2] Exclude “show-and-tell” cases in which the original turn content already had the answer and wanted to show examples. In this case, replies will not be informative.

username	content	label
柯**所	患者 女 22岁 双大腿外侧冷球蛋白血症性股臀部皮肤血管炎7年[如图]. 处理:冻疮膏+倍他米松新霉素乳膏 交替用 效佳 [本帖最后由 tiandou1975 于 2008-2-5 15:07 编辑]	EXCLUDE_THREAD
x****7	很好，学习了。:victory:	
n****i	我一般用派瑞松:lol: 与倍他米松新霉素乳膏机理相同 [本帖最后由 tiandou1975 于 2008-2-6 07:21 编辑]	
a****2	恩，抗组织胺药物也应有效。 多谢楼主分享。 [本帖最后由 adcc102 于 2008-2-5 13:14 编辑]	
o****6	应诊断：股臀部皮肤血管炎！ 赞同你的诊断 [本帖最后由 tiandou1975 于 2008-2-6 07:22 编辑]	

[3] Exclude cases with no medical context content in the original turn. In this case neither the query or the replies will be informative.

username	content	label
x****i	谢谢	EXCLUDE_THREAD
哈***V	这个照片，真正的雾里看花！ 1.丘疹性荨麻疹？ 2.毛囊炎？ 3.传染性软疣？	
超*	丘疹性荨麻疹？	
y****k	不是很看的清楚	
d****0	丘疹性荨麻疹	

III. Turn Labeling and Exclusion

The turn exclusion stage refers to cases in which the conversation has not been excluded from the prior step, but individual turns require removal. These labels are used as helper labels for the content editing and automatic query-response construction stages.

Table 4. Turn labels

label	description	examples
valid_query	Turns written by the original author that includes medical information or medical questions. Example of non medical information or medical questions: [“thanks!”,“anyone there?”] A conversation may have multiple valid_query labels, however in the content-editing stage, a combined version is created.	是荨麻疹还是啥的 十年了有了慢慢全身都是
profadv	Turn content that includes professional advice Example: diagnosis, test, or treatment suggestions If the turn content quotes other turns with professional advice and agrees, this is also allowed.	寻常狼疮 查一下过敏原吧
reqclinicalinfo	Request additional information about history of present illness with NO professional advise Examples: questions about the current problem or questions about clinical history or patient demographics	图片模糊 去做过“过敏原点刺实验”吗？ 其他地方有吗？
reqclinicalinfo_and_profadv	Combination of the two above labels	图片模糊，初步考虑丘疹性荨麻疹。 去做过“过敏原点刺实验”吗？查一下过敏原吧 其他地方有吗？怎么有点像匍行诊呀 是起了就消失不了吗？还是时起时消，如果是后者考虑是荨麻疹，如果是前者考虑是螨虫性皮炎。
exclude_post	Turn content that does not offer additional useful information. This can originate from the original conversation author or from replies. This label may also be used if the turn content cannot be disentangled from the rest of the conversation through content editing.	谢谢！ 期待结果

IV. Turn Content-Editing

Context-free and Metadata Cleaning Edits

Text content that references other turn content should be edited to be independent. Furthermore, automatically generated time stamps or signatures should be removed.

Table 5. Examples of metadata to remove

Edit_scenario	Description	Example	content_edited
Make text context-free	Edit to remove references to other examples	支持血管痣	血管痣
	If quoting another turn to voice approval, remove the quote and keep the content.	“血管痣” +1	血管痣
Website name (爱爱医) appears as they appear in text	Edit away references to the iiyi website	请爱爱医的医生帮我看下	请医生帮我看下
Autogenerated metadata	Edit away automatically generated text	患儿手部有多数椭园形小水疱 [本帖最后由 23637358 于 2008-6-25 22:03 编辑]	患儿手部有多数椭园形小水疱

Remove “Updated Information” Edits

In some cases, authors may update their original content with the answer (these are typical of doctor-initiated conversations). In order to conserve the cases, we edit to remove these updated answers.

You can tell if the turn content had the answer originally by (a) an update time-stamp (b) by the responses of other authors, (c) and the language of the turn content.

Table 6. Examples of query turn content (from different threads) updated to include the answer (公布答案 花斑癣). In these cases, when reading the entire conversation, it is clear the answer was not originally in the turn content. For these, we can edit the content to remove the later added answer.

Original Turn Content	Edited Original Turn Content
患者男性，45岁。躯干皮损时轻时重，已7年了，。请大家给个诊断 公布答案 花斑癣 [本帖最后由 23637358 于 2008-5-29 10:14 编辑]	患者男性，45岁。躯干皮损时轻时重，已7年了，。请大家给个诊断
患儿手部有多数椭园形小水疱 公布答案:手足口病	患儿手部有多数椭园形小水疱

Combine Valid Queries, Leave Empty Content-Edited Turns For Unpaired Query/Responses

For multiple valid_query turns, a combined final query should be edited and placed in the last relevant valid_query label turn. All previous valid_query turns should be left empty. Responses that occur prior to query content should have content_edited left blank (or marked with exclude_post). In the same vein, responses that require query content information from a subsequent query should have content_edited left blank (or marked with exclude_post).

For example: In the below table, response from Turn 2, only has input information from Turn 1; whereas responses from 4-6 have access to all the patient information from both Turn 1 and 3. Likewise, responses from Turns 8-9 have access to Turns 1, 3, and 7.

Table 7. Example of query-response sets

turn	username	query_example	input
1	thread_author	What is this? Started last week.	
2	p1		Turn 1
3	thread_author	21 years old. No other allergies.	
4	p2		Turn 1, 3
5	p3		Turn 1, 3
6	p4		Turn 1, 3
7	thread_author	I tried antihistamine and it didn't work.	

8	p4		Turn 1, 3, 7
9	p5		Turn 1, 3, 7

Thus, three possible valid query and response combination sets are valid:

- Option 1:
    - content_edited for Turns 1-2 are non-empty
    - Turn 1 content_edited becomes the query
        - **Query**: “What is this? Started last week.”
    - Turn 2 content_edited becomes the only response
    - Turns 3-9 content_edited should be left blank.
- Option 2:
    - content_edited for Turns 3-6 are non-empty
    - Turns 3 content_edited will include combined information from the original Turn 1 and 3 content.
        - **Query**: “What is this? Started last week. 21 years old. No other allergies.”
    - Turns 4-6 content_edited becomes individual responses for a total of 3 responses.
    - Turns 1-2, 7-9 content_edited should be left blank.
- Option 3:
    - content_edited for Turns 7-9 are non-empty
    - Turn 7 content_edited will include combined information from the original Turns 1, 3, and 7 content.
        - **Query**: “What is this? Started last week. 21 years old. No other allergies. I tried antihistamine and it didn’t work.”
    - Turns 8-9 content_edited becomes individual responses for a total of 2 responses.
    - Turns 1-6 content_edited should be left blank.

The annotator should identify the best combined query to edit based on the algorithm which will lead to the most responses. Thus, in this case, Option 2 which gives 3 responses will be the correct editing strategy. (See Section VI Detailed Annotation Questions 3 for another detailed example).

V. Automatic Query-Response Construction

The automatic query-response construction will involve five steps:

1) Exclude conversations with **EXCLUDE_THREAD** label
2) Remove all turns with labels **exclude_post** and **reqclinicalinfo**
3) Remove all empty content_edited turns
4) Concatenate remaining consecutive content_edited turns by the same author.
5) Take the first non-empty content_edited valid_query as the textual **query**, remove all other turns by the original author
6) Each remaining content_edited text turn, not authored by the original conversation author, will become a **response**.

Below is an example of the algorithm at work given the labeled turns and content_edited text.

Table 8. Example of final threads kept using the content_edited column

username	content	label	content_edited	kept	keep_reason
original_author	content1	valid_query		0	Excluded, no content
replier1	content2	profadv	content2-edited	0	Excluded, first turn should be by original author
original_author	content3	valid_query	content1+content 3-edited	1	Kept, first non-empty content_edited by original author
replier2	content4	reqclininfo	content4-edited	0	Excluded, not an advice label
replier3	content5	profadv	content5-edited	1	Kept, non-empty, non-excluded label
original_author	content6	exclude_post		0	Excluded, empty content_edited
replier4	content7	reqclinicalinfo_ and_profadv	content7-edited	1	Kept, non-empty, non-excluded label

This is converted to a query-response structure such as:

```
{
      “query”: content1+content3-edited,
      “author”: original_author,
      “responses”: [
            {     “author”:   replier3,
                  “content”:    content5-edited
            },
            {     “author”:   replier4,
                  “content”:    content7-edited
            },
    ]
}
```

VI. Detailed Annotation Questions

1. Merging valid queries (from the original author)

content	label	content_edited
去年8月吧，在脸颊两边一群痘痘一夜之间冒出来，到现在还没好象还有痘痕	valid_query	
详细说明年龄性别和发病及治疗情况！	reqclinicalinfo	
这是我的图片~几个月不消去 21岁 我重新编辑一下您的图片并发出，但图片的像素似乎比较低，可[本帖最后由 鱼er 于 2006-2-14 13:03 编辑]	valid_query	
图片太模糊了~~~	reqclinicalinfo	
还模糊啊，颜色红的就是痘	valid_query	21岁，去年8月吧，在脸颊两边一群痘痘一夜之间冒出来，到现在还没有好，有恶化的趋向。。好象还有痘痕
市面上的蓝金组合还不错，不过挺贵的，我们这老大夫用甲氰吩	profadv	市面上的蓝金组合还不错，不过挺贵的，我们这老大夫用甲氰咪胍1片一天3次和葡萄糖酸辛口服一月用听说

2. If the original author asks a follow-up question later, and there are no (or less replies if counting from the later valid_query), then choose to exclude the last turn from the original author.

1	username	content	label
2	p****冬	各位医生大夫: 你们好,我是从事IT行业的,主要是靠	valid_...
3	蓝*	可以行微波手术,但要注意烧彻底/	profadv
4	p****冬	去哪里可以做,需要多少钱呀	exclud...
5	蓝*	一般大医院皮肤科都可以做/	exclud...
6	x****g	我刚发了个治疗鸡眼的录像，有配音解说 https://bbs.iiyi.com/forum.php?m ... &ext 蓝风版主：你好，我本来应该把它发在皮	exclud...
7	蓝*	好,谢谢分享~	exclud...
8	辽*人	鸡眼是一种物理型疾病,不能排除 但看到您的图片和所述症状寻常疣的可能 建议停用鸡眼膏观察	profadv
9	y****n	寻常疣吧，建议能冻，有的也可刮除，有	profadv
10	林*玉	应该是寻常疣吧～～这个地方长鸡眼实在	profadv
11	滑*	我刚发了个治疗鸡眼的录像，有配音解说 我下载了 怎么不能看	exclud...
12	x****g	我再说明一次吧：用暴风影音可以打开	exclud...
13	p****冬		exclud...
14	x****g	原帖由 滑翔 于 2006-7-9 08:45 发表 我刚发了个治疗鸡眼的录像，有配音解说 我下载了 怎么不能看 你看帖子仔细点，包括回复贴； 不会解压的爱友：麻烦你检查自己的下载	exclud...
15	★****★	鸡眼面不大可能,应该考虑寻常疣,可以冷 还有就是切处,不过不管用什么治疗都可能	profadv
16	p****冬	常疣是什么样的,可否解释一下	exclud...

3. Multiple query-response pattern alternatives

If multiple valid queries from the original author exist, it is possible to have multiple patterns of query-response exclusion or inclusion. In this case, the guidance on selecting between alternatives is to maximize patterns that keep the largest number of professional advice replies.

If deciding between alternative cases:

1. Try to keep as much background information and professional advice as possible
2. If the original authors' have multiple turn content in the first few turns, we tend to keep them and concatenate them
3. If the original author's new turns are in the last few turns of the conversation, we tend to exclude them, even though those turns may contain new information.

<table>
<tr><th>Label description</th><th>Example turn content editing</th></tr>
<tr><td>Option1:
Keep first valid_post and discard everything at second valid_post and after.

[NOT OPTIMAL]</td><td>
<table>
<tr><th>turn</th><th>username</th><th>content</th><th>label</th><th>content_edited</th></tr>
<tr><td>1</td><td>p1</td><td>这是什么病？</td><td>valid_query</td><td>这是什么病？</td></tr>
<tr><td>2</td><td>p2</td><td>什么时候开始的？是不是血管瘤?</td><td>reqclinicalinfo_and_profadv</td><td>什么时候开始的？是不是血管瘤?</td></tr>
<tr><td>3</td><td>p1</td><td>三个星期前开始的。本人21岁</td><td>valid_post</td><td></td></tr>
<tr><td>4</td><td>p3</td><td>是不是血管瘤。蜘蛛痣</td><td>profadv</td><td></td></tr>
<tr><td>5</td><td>p4</td><td>支持血管痣</td><td>profadv</td><td></td></tr>
</table>
</td></tr>
<tr><td>Option2:
Keep the concatenated p1 turn into second valid_post turn, and discard everything the first reqclinicalinfo_and_profadv.

[OPTIMAL]</td><td>
<table>
<tr><th>turn</th><th>username</th><th>content</th><th>label</th><th>content_edited</th></tr>
<tr><td>1</td><td>p1</td><td>这是什么病？</td><td>valid_query</td><td></td></tr>
<tr><td>2</td><td>p2</td><td>什么时候开始的？是不是血管瘤?</td><td>reqclinicalinfo_and_profadv</td><td></td></tr>
<tr><td>3</td><td>p1</td><td>三个星期前开始的。本人21岁</td><td>valid_query</td><td>这是什么病？三个星期前开始的。本人21岁</td></tr>
<tr><td>4</td><td>p3</td><td>是不是血管瘤。蜘蛛痣</td><td>profadv</td><td>是不是血管瘤。蜘蛛痣</td></tr>
<tr><td>5</td><td>p4</td><td>支持血管痣</td><td>profadv</td><td>血管痣</td></tr>
</table>
</td></tr>
</table>