Evaluation guideline for dialogue summarization

Your task is to evaluate summaries generated by different systems as part of a dialogue summarization task. You have to read the original dialogue, then rate each summary on various criteria using a likert scale from 1 to 5, with **5 being the best value for a criterion**.

Please read this guide carefully and in full before starting the evaluation task on the platform.

1. Introduction about the task

Automatic dialogue summarization aims to identify the most important content from human conversations to synthesize it in a short textual summary. In this task, the dialogues are the ones extracted from a call-center of a public transport entity and which deals mainly with customer inquiries and agent responses.
To evaluate the ability of different systems to produce a short summary from a dialogue, your task will be to evaluate 6 summaries on some criteria using a likert scale from 1 to 5.

2. Evaluation platform

You will be presented to an interface as follow[1]:

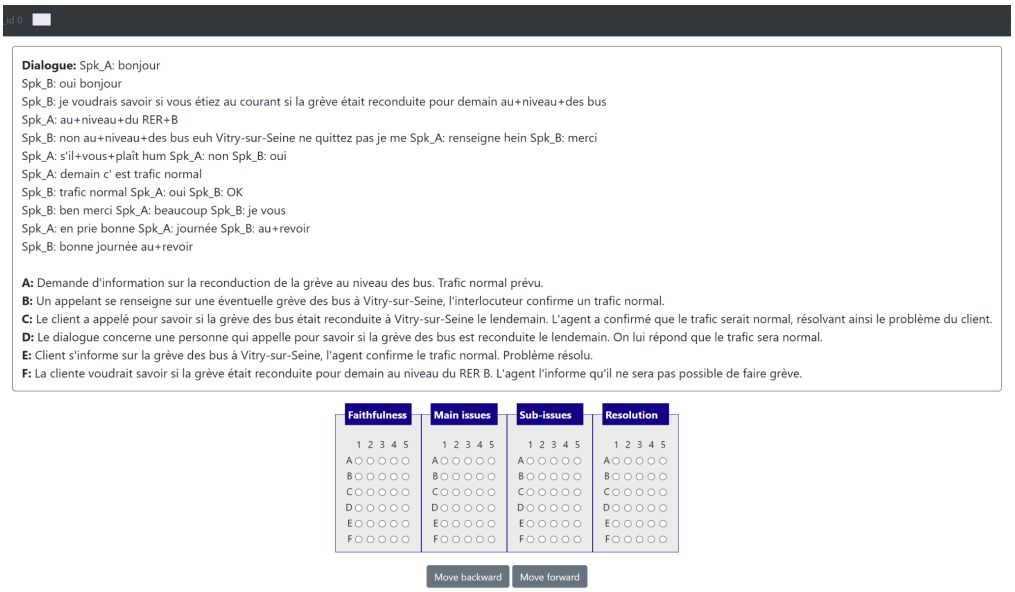

Figure 1: Interface of evaluation platform.

[1] A link will be provided: *(anonymised link)*, replace "user" at the end with your first name, and you'll be redirected to the interface

The evaluation interface, shown in Figure 1, features a dialogue at the top, accompanied by six different summaries (named A, B, C, D, E, F).

3. Evaluation Task Instructions

Understanding the Interface: Refer to Figure 1, which illustrates 4 blocks representing different evaluation criteria. Behind the blocks, click on the “Move forward” button to move on to the next annotation, and “Move backward” to return to the previous annotation.
Evaluation Process: Within each block, interact with the interface by clicking on each of the six summaries. Assign a score from **1 to 5 (the higher the better)** to each summary based on its adherence to the specific criterion.
Submission: Once you have evaluated all summaries within each block, proceed to the next block until you have completed the evaluation for all criteria.
Completion: Ensure that you have provided **a score for each summary in every block** before finalizing your evaluation. Click on the “Move forward” button behind the blocks to move on to the next annotation. When you have completed the evaluation of all dialogue-summaries pairs and you see “***Finished 20/20***” in the top left-hand corner, the task is complete.

The descriptions of each criterion are explained below:

1) **Faithfulness**
   Le résumé doit respecter le dialogue au niveau des informations factuelles. Les niveaux de Likert en 5 points sont les suivants :
   - **1 : Très peu fidèle :** Le résumé contient de nombreuses erreurs factuelles et omet des informations importantes du dialogue.
   - **2 : Peu fidèle :** Le résumé contient quelques erreurs factuelles et/ou omet certaines informations importantes du dialogue.
   - **3 : Moyennement fidèle :** Le résumé est globalement cohérent avec le dialogue, mais contient quelques petites erreurs ou inexactitudes.
   - **4 : Assez fidèle :** Le résumé est très proche du dialogue, avec très peu d'erreurs factuelles ou d'inexactitudes mineures.
   - **5 : Très fidèle :** Le résumé est parfaitement cohérent avec le dialogue, sans erreur factuelle ni omission importante.
2) **Main issues**
   Dans le résumé, il faut noter les sujets principaux de la conversation. C’est-à-dire les problèmes pour lesquels le client a appelé. L’identification de ces problèmes constitue la base de la classification de l'appel dans plusieurs classes différentes de motivations d'appel. Le problème principal d'un appel doit être classé par ordre de priorité pour être inclus dans le résumé du même appel.
   Les niveaux de Likert en 5 points sont les suivants :
   - **1 : erroné ou manquant :** Les questions principales de l'appel sont présentées de manière incorrecte, sont manquantes ou sont complètement différentes de celles qui ont été discutées pendant l'appel.
   - **2 : confus :** Les questions principales de l'appel sont présentées de manière confuse ou ambiguë, ce qui rend difficile la compréhension des problèmes abordés pendant l'appel.

- **3 : correct mais incomplet :** Les questions principales de l'appel sont correctement identifiées, mais la présentation manque de détails ou de contexte pour comprendre pleinement les problèmes abordés.
- **4 : correct mais pourrait être amélioré :** Les questions principales de l'appel sont correctement identifiées et présentées de manière claire, mais il manque quelques détails ou précisions pour une compréhension complète.
- **5 : précis et synthétique :** Les questions principales de l'appel sont présentées de manière claire, concise et précise, permettant une compréhension complète des problèmes abordés pendant l'appel.

3) **Sub-issues**

Dans le résumé, il faut également noter les sous-problèmes de la conversation : lorsqu'un sous-problème apparaît dans la conversation, il peut être introduit par le client ou par les agents.

Les niveaux de Likert en 5 points sont les suivants :

- **1 : erroné ou manquant :** Les sous-problèmes de l'appel sont présentés de manière incorrecte, sont manquantes ou sont complètement différentes de celles qui ont été discutées pendant l'appel.
- **2 : confus :** Les sous-problèmes de l'appel sont présentés de manière confuse ou ambiguë, ce qui rend difficile la compréhension des problèmes abordés pendant l'appel.
- **3 : incomplet :** Les sous-problèmes de l'appel sont correctement identifiés, mais la présentation manque de détails ou de contexte pour comprendre pleinement les problèmes abordés.
- **4 : correct mais pourrait être amélioré :** Les sous-problèmes de l'appel sont correctement identifiés et présentés de manière claire, mais il manque quelques détails ou précisions pour une compréhension complète.
- **5 : précis et synthétique :** Les sous-problèmes de l'appel sont présentés de manière claire, concise et précise, permettant une compréhension complète des problèmes abordés pendant l'appel.

Note : Si aucun sous-problème n'est présenté dans la conversation, laissez le bloc « Sub-issues » vide pour tous les résumés.

4) **Resolution**

Dans le résumé, il faut également noter la résolution de l'appel : c'est-à-dire si le problème du client a été résolu au cours de cet appel (résolution au premier appel) ou non. Les niveaux de Likert en 5 points sont les suivants :

- **1 : absent :** Le résumé ne mentionne pas si le problème du client a été résolu ou non.
- **2 : confus :** Le résumé mentionne vaguement la résolution, mais il est difficile de comprendre si le problème a été résolu ou non.
- **3 : présent mais incomplet :** Le résumé mentionne que le problème a été résolu, mais les détails sont manquants ou la résolution n'est pas présentée de manière claire.
- **4 : correct mais pourrait être amélioré :** Le résumé mentionne que le problème a été résolu et fournit quelques détails, mais il manque encore des informations pour avoir une compréhension complète.
- **5 : précis et synthétique :** Le résumé présente clairement et de manière détaillée la résolution du problème du client, permettant de comprendre comment le problème a été résolu.

Voici un exemple :

Dialogue :

<Spk_A> bonjour
<Spk_B> allo oui bonjour
<Spk_B> euh je me permettais en+fait de vous appeler pour un renseignement renseignement alors j' ai mon mari qui travaille à la RATP
<Spk_A> hm+hm
<Spk_B> et je voulais savoir en+fait euh est-ce+que via son travail est-ce+qu' on
<Spk_B> peut bénéficier d' une réduction par+exemple pour sa femme
<Spk_A> oui bien+sûr
<Spk_B> oui
<Spk_A> il faut qu' il fasse la demande auprès+de son service administratif
<Spk_B> d'accord
<Spk_A> il aura une carte de famille euh d' agent qui permet d' avoir euh cinquante pourcent de réduction sur les titres de transport <Spk_A> enfin sur les titres de transport <Spk_B> ah d'accord très+bien
<Spk_A> uniquement sur le carnet hein <Spk_A> les billets hein <Spk_B> d'accord très+bien
<Spk_A> pas sur les cartes Orange ou les choses comme+ça
<Spk_B> d'accord très+bien
<Spk_B> merci beaucoup
<Spk_A> je vous en prie au+revoir <Spk_A> madame <Spk_B> au+revoir

Exemple de résumé 1 :
Une femme appelle pour savoir si elle peut bénéficier d'une réduction sur les titres de transport de la RATP grâce au travail de son mari. On lui explique la procédure à suivre.

- **Faithfulness** : 4 (Le résumé est globalement factuellement correct, mais il pourrait être amélioré en fournissant plus de détails sur ce que l'agent a expliqué à la cliente)
- **Main issues** : 5 (La question principale est clairement présentée et bien résumée)
- **Sub-issues** : N/A (Il n'y a pas de sous-problème dans cette conversation)
- **Resolution** : 3 (L'agent a répondu à la question, "On lui explique la procédure à suivre", mais la résolution n'est pas présentée en détail)

Exemple de résumé 2 :
Demande de renseignement sur l'obtention d'une réduction sur les titres de transport. Communication du numéro du service concerné.

- **Faithfulness** : 2 (Le résumé contient des informations factuellement incorrectes, notamment la mention de la "Communication du numéro du service concerné" qui n'a pas eu lieu dans l'appel)
- **Main issues** : 4 (La question principale est présentée et résumée mais il n'y a pas d'information sur le fait que le mari de cette femme travaille à la RATP)
- **Sub-issues** : N/A (Il n'y a pas de sous-problème dans cette conversation)
- **Resolution** : 2 (La résolution est incomplète, car le résumé ne mentionne pas les informations fournies par l'agent sur la procédure à suivre, le taux de réduction et son application aux différents types de titres de transport)