Human Evaluation: Quality of Argument Summaries

This task is about *Evaluating the Quality of Argument Summaries*. To this end, you will be confronted with two subtasks. Before starting with the description of these subtasks, the concept of *Argument Summarization* will be briefly explained.

Argument Summarization

A debate on a certain topic can be conducted using a variety of arguments for each side of the debate. Although some of these arguments refer to the same main statement, they can be formulated very differently. While the number of possible arguments seems to be almost infinite due to the possibility of different formulations, the number of possible main statements in a debate is limited.
Argument summarization is about summarizing a relatively large set of arguments on a certain debate topic and stance by generating a small set of argument summaries, each expressing one distinct main statement contained in the set of arguments. In addition, each argument is matched to the generated summary that conveys its main statement the best. Following is a simple example:

Topic: We should abandon the use of school uniform

Stance: Opposing

Set of Arguments:

1. School uniforms keep everyone looking the same and prevent bullying.
2. School uniforms can help parents save money on outfits.
3. School uniforms help stop bullying because when people are similarly dressed, nobody is made to feel inferior.
4. It is cheaper for parents to buy school uniforms, which is helpful to parents that are struggling financially.
5. School uniforms are substantially more affordable.

Set of Summaries:

1. School uniforms reduce bullying.
2. School uniforms save costs.

Argument Summary Matches:

The matches are highlighted by the colored markings:

- Arguments 1 and 3 are matched to summary 1
- Arguments 2, 4 and 5 are matched to summary 2

Evaluating the Quality of Argument Summaries

Subtask 1: Quality of Summaries (reference-based)

This subtask is about determining how well a set of generated summaries serve as a summary of possible arguments for the topic and stance at hand.

For this purpose, you are given a set of generated summaries and a set of reference summaries as well as the corresponding debate topic and stance. You have to carry out the following two instructions regarding the criteria of coverage and uniqueness:

1. **Coverage**: Count the number of reference summaries that are covered by the set of generated summaries.
2. **Uniqueness**: Count the number of distinct/unique main statements contained in the set of generated summaries.

For both criteria increments of 0.5 are allowed. In the case of coverage, this applies if a reference summary is only partially covered by the set of generated summaries. For the criterion of redundancy, this applies if there is a distinct main statement in the set of generated summaries that partially overlaps with another. For the case you are not sure, you can answer with -1. Following is an example:

Topic: Routine child vaccinations should be mandatory

Stance: Opposing

Set of reference summaries:

1: Mandatory vaccination contradicts basic rights
2: Routine child vaccinations are not necessary to keep children healthy
3: Routine child vaccinations, or their side effects, are dangerous
4: The parents and not the state should decide

Set of generated summaries:

1: Vaccinations violate free will and personal choice.
2: Mandatory vaccines conflict with religious beliefs
3: Parents should have the right to decide.
4: Children may suffer harmful effects from vaccines.
5: Concerns about vaccine safety and side effects.

Coverage: 3

Explanation**:** The second reference summary is not covered.

Uniqueness: 3.5

Explanation: The first and third generated summaries address two different distinct main statements. The fourth and fifth generated summaries refer to the same distinct main statement. The second generated summary partially overlaps with the first one.

Evaluation Task 2: Quality of Argument Summary Matches (reference-free)

This evaluation task will follow.