HUMAN EVALUATION

A convincing evaluation of a summarization system requires human judgement. In this study, we ask participants to rate summaries from aspects of faithfulness, informativeness, readability, and conciseness. Instructions with examples are presented in the following. Each participant needs to 1) complete a qualification test, which is used to estimate the confidence level of the judgements; 2) evaluate 50 samples that we provide.

INSTRUCTIONS

We illustrate the features: faithfulness, informativeness, readability, and conciseness.

1) **Faithfulness.** Faithfulness means factual consistency with the context. Please avoid using general knowledge, and only consider it in the context of the provided document. The summary is inconsistent if facts in the summary are not supported by the document. Two typical cases are conflict and hallucination.

(i) The summary contradicts the information in the document. The summary might say ”A fire broke out in Seattle”, but the document says it broke out in Portland. Or the summary might say ”the Republicans won the election”, but the document indicates the Democrats won instead.

(ii) The summary adds (hallucinates) a fact that is not mentioned anywhere in the document. For example, the summary might say that ”A fire broke out at 2 am”, but the document does not mention the time when the fire broke out.

2) **Informativeness.** It means that a summary expresses the main points of the document. A summary should contain relevant and important information and few unimportant details. If you select the summary to be not consistent with the document, please only consider the consistent information when evaluating this category.

3) **Readability.** The summaries are written by human or generated by language models. A summary is readable/fluent if free from language problems. A less readable summary is confusing and difficult to understand.

4) **Conciseness.** It means that a summary contains little *irrelevant or minor* information. A summary should capture important information and avoid excessive redundancy. Please just consider less relevant information. And concise does not mean short.

We present an example in Table 1. This article reports an accidental death of Alexys Brown. The main information is the accident and the appeal to raise money, and minor points can be the investigation, post-mortem examination, etc. Summary 1 can be a reasonable and acceptable summary. Summary 2 misses a major point, the appeal, thus not informative. Both summary 3 and summary 4 shows unfaithfulness. Summary 3 makes a factual mistake that Alexys died *of cancer*. Thus contradicts the article. Summary 4 adds a hallucination that Alexys is *three-year-old*. Summary 5 is confusing because grammar flaws impair readability. Summary 6 covers main points but gives some unimportant information like what Alison said, thus not concise.

QUALIFICATION TEST

You are given 5 articles and 2 summaries (numbered as #1, #2) for each article. Please first read the article and then evaluate the quality of the summaries from the four aspects that are introduced above, i.e faithfulness, informativeness, readability, and conciseness.

Please annotate which one of the 2 summaries is better in the four aspects separately. For example, if summary #1 is better than summary #2 in the aspect of informativeness, then type “1” behind ‘*****Informativeness:‘. It means summary #1 wins over summary #2 on informativeness. And if summary #2 wins in readability, then type “2” behind ‘*****Readability:‘. If the two summaries draw with each other (come out even) in an aspect, then type “0” in the cell below that aspect.

Your results are 4 integers for each examples. The scores are in the order of *faithfulness, informativeness, readability, conciseness*. The format is shown in Table 2.

EVALUATION

You are given 50 articles, and 2 summaries (numbered as #1, #2) for each article. Please read the article and then evaluate the quality of the summaries in the same way as the qualification test. Your results are 4 integers for each examples. The scores are in the order of *faithfulness, informativeness, readability, conciseness*. The format is shown in Table 2. There may be repeated summaries of an article, and please make sure that their scores are all 0.

Article: Alexys Brown, also known as Lexi, died at her home in Emmadale Close, Weymouth, on Thursday. An investigation is under way to discover how she became trapped. A post-mortem examination is due to be carried out this week. It was originally hoped the appeal would raise £2,000. Alison Record, who started the Just Giving appeal, said she was ”heart broken” over the death. “Everybody by now has heard of the terrible tragedy the Brown family have suffered with the loss of their beautiful and beloved little girl Lexi,” the appeal page reads. Many other comments have been posted on the appeal page. Steph Harris said: “Thinking of you all at this devastating time, fly high beautiful princess. Love Steph and family xxx” Lesley Andrews added: “No amount of money will take away the pain, but so much love comes with every penny. Take care. xx” Aster Group, the housing association responsible for managing the home, is assisting with the police investigation. The Health and Safety Executive (HSE) is also investigating. Dorset County Council said it had not installed the disabled lift at the property.
Summary #1: An appeal to raise money for the family of a girl who died after getting stuck in a lift was originally hoped for raising 2,000 pounds.
Summary #2 (informativeness): Alexys Brown, also known as Lexi, died at her home in Emmadale Close, Weymouth, on Thursday.
Summary #3 (faithfulness): Alexys Brown, also known as Lexi, died of cancer. The appeal was originally hoped for raising 2,000 pounds.
Summary #4 (faithfulness): An appeal to raise money for Alexys Brown, a three-year-old girl who died after getting stuck in a lift was originally hoped for raising 2,000 pounds.
Summary #5 (readability): An appeal to raise the family of Alexys Brown became trapped in a lift would raise 2,000 pounds.
Summary #6 (conciseness): An appeal to raise money for Alexys Brown, who died after getting stuck in a lift was originally hoped for raising 2,000 pounds. It was originally hoped the appeal would raise £2,000. Alison Record, who started the Just Giving appeal, said she was ”heart broken” over the death. “Everybody by now has heard of the terrible tragedy the Brown family have suffered with the loss of their beautiful and beloved little girl Lexi,” the appeal page reads.
Summary #7: Alexys Brown, also known as Lexi, died from cancer in Emmadale Close, Weymouth, on Monday. An investigation is under way under way to discover how she became trapped.

Table 1: An example for illustration.

Article:
Alexys Brown, also known as Lexi, died at her home in Emmadale Close, Weymouth, on Thursday. An investigation is under way to discover how she became trapped. A post-mortem examination is due to be carried out this week. It was originally hoped the appeal would raise £2,000. Alison Record, who started the Just Giving appeal, said she was ”heart broken” over the death. “Everybody by now has heard of the terrible tragedy the Brown family have suffered with the loss of their beautiful and beloved little girl Lexi,” the appeal page reads. Many other comments have been posted on the appeal page. Steph Harris said: “Thinking of you all at this devastating time, fly high beautiful princess. Love Steph and family xxx” Lesley Andrews added: “No amount of money will take away the pain, but so much love comes with every penny. Take care. xx” Aster Group, the housing association responsible for managing the home, is assisting with the police investigation. The Health and Safety Executive (HSE) is also investigating. Dorset County Council said it had not installed the disabled lift at the property.

Summary #1:
Alexys Brown, also known as Lexi, died at her home in Emmadale Close, Weymouth, on Thursday.

Summary #2:
Alexys Brown, also known as Lexi, died of cancer. The appeal was originally hoped for raising 2,000 pounds.

Scores:
*****Faithfulness: 2
*****Informativeness: 2
*****Readability: 0
*****Conciseness: 2

Table 2: An example for the format of the scores.