Evaluation Guidelines for Human Judges (Final Version)

The purpose of this evaluation is to assess questions that have been generated from simple input sentences.

Three pieces of information will be provided for each evaluation:

1) The input sentence.

2) The requested question type for the generated question. E.g. who, what, where, when, etc.

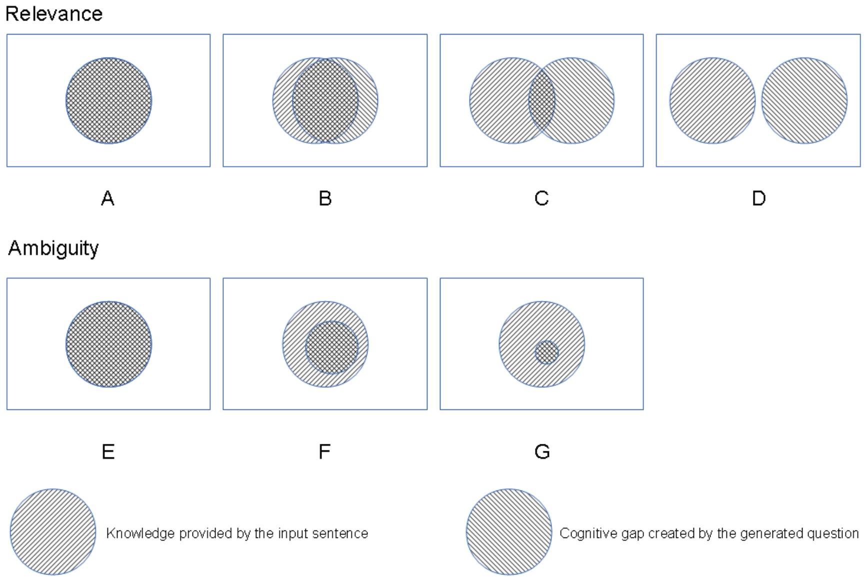
3) The generated question.

The generated questions are to be evaluated using the five criteria: Relevance, Question type, Correctness, Ambiguity, and Variety. **It must be stressed that these criteria are to be applied independently of each other.**

Each criteria, defined below, is assigned a rank, with **1** being the **greatest**.

Language is complex and not all evaluations will be so clear cut as the examples given below. When dealing with such an evaluation it can be helpful to think of it as a puzzle. A sensible approach would be to eliminate any obviously incorrect-ranked criteria first and then attempt to make an argument for the remaining criteria. To assist with this approach, where appropriate, each of the criteria below identifies a key point to consider when making the evaluation.

Relevance and ambiguity are more difficult to calculate. The following visual representation is helpful:


Relevance
A
B
C
D
Ambiguity
E
F
G
Knowledge provided by the input sentence
Cognitive gap created by the generated question

A-D represent Relevance ranks 1-4 respectively, and E-F represent Ambiguity ranks 1-3 respectively.

RELEVANCE

Questions should be relevant to the input sentence. This criterion measures how well the question can be answered based on what the input sentence says.

One consideration when assessing this criteria is the following: Only the information provided in the input sentence is important. **A question, where the answer cannot be found in the input sentence is not relevant.**

Example input sentence: John likes apples and pears.

Rank	Description	Example
1	The question is completely relevant to the input sentence.	What fruit does John like?
2	The question relates mostly to the input sentence.	Does John like apples and oranges?
3	The question is only slightly related to the input sentence.	Does John like all fruit?
4	The question is totally unrelated to the input sentence.	Does John like oranges?

QUESTION TYPE

Questions should be of the specified target question type.

Example target question type: **What**

Rank	Description	Example
1	The question is of the target question type.	**What** fruit does John like?
2	The type of the generated question and the target question type are different.	**Why** does John like apples?

SYNTACTIC CORRECTNESS AND FLUENCY

The syntactic correctness is rated to ensure systems can generate sensible output. In addition, those questions which read fluently are ranked higher.

One consideration when assessing this criteria is to ask the following question: **Can the question be answered?** If you cannot work out what the answer is supposed to be then the criteria must be rank 4.

Rank	Description	Example
1	The question is grammatically correct and idiomatic/natural.	In which type of animals are phagocytes highly developed?
2	The question is grammatically correct but does not read as fluently as we would like.	In which type of animals are phagocytes, which are important throughout the animal kingdom, highly developed?
3	There are some grammatical errors in the question.	In which type of animals is phagocytes, which are important throughout the animal kingdom, highly developed?
4	The question is grammatically unacceptable.	On which type of animals is phagocytes, which are important throughout the animal kingdom, developed?

AMBIGUITY

The question should make sense when asked more or less out of the blue. Typically, an unambiguous question will have one clear answer.

One consideration when assessing this criteria is to ask the following question: **Can more information be added from the input sentence to make the question more specific?** If so then the question has some ambiguity. Similarly, if information can be removed from the question then the question is not completely ambiguous.

Example input sentence: Sonia Maria Sotomayor was nominated in 1997 to the U.S. Court of Appeals for the Second Circuit.

Rank	Description	Example
1	The question is unambiguous.	Who was nominated in 1997 to the U.S. Court of Appeals for the Second Circuit?
2	The question could provide more information.	Who was nominated in 1997?
3	The question is clearly ambiguous when asked out of the blue.	Who was nominated?

VARIETY

Pairs of questions in answer to a single input are evaluated on how different they are from each other. One consideration when assessing this criteria is to ask the following question: **are the answers to both questions different?** If so then no matter how closely worded the questions are, they have a rank 1 variety.

Example input sentence: John was born in London but worked in Glasgow.

Rank	Description	Example
1	The two questions are different in content.	Where was John born? Where did John work?
2	Both ask the same question, but there are grammatical and/or lexical differences.	Where was John born? What was John's place of birth?
3	The two questions are identical.	Where was John born? Where was John born?