Human Annotation Guidelines for Automatic Text Generation

1 For Machine Translation task

Guidelines -

The following guidelines are for the human evaluation of the automatic neural machine translation task,

- Use the Romanian or the German as the source data and the English as the target data in the given excel file.
- English sentences are the machine translated data which needs to be evaluated with human annotation scores.
- The translations should be evaluated sentence wise. Each sentence should have an individual annotation score irrespective of other sentences.
- The human annotation score should be for the semantics of the translated sentence and not for lexical(similar words) translation.
- The translations should be non-redundant and informative.
- The annotator should keep in mind about both the fluency and the adequacy of the translated sentences and give one score for each sentence.
- The annotation scores are in the range of 1-5 and the details of each score is in the table below

Score	Meaning	Details
1	Very bad	The (EN) translation has (very) different meaning compared to the original (RO/DE) text
2	Bad	The (EN) translation has some meaning overlap with the original text, but is mostly incorrect
3	Borderline	The (EN) translation roughly corresponds to the original (RO/DE) text, but there are non-negligible translation errors
4	Good	The (EN) captures well the meaning of the original (RO/DE) text, with some minor (generally negligible) errors
5	Very good	The (EN) text is an accurate translation of the original (RO/DE) text

2 For Text Summarization task

2.1 Generates short text summaries not just the title

Guidelines -

The following guidelines are for the human evaluation of the automatic abstractive text summarization task,

- Use the English long sentences/paragraphs in the first column as the source data and the short English summaries in the second column as the target data in the given excel file.
- The short summaries are the machine summarized text which needs to be evaluated with human annotation scores.
- In case of models generating titles only , see subsection 2.2 .
- The summarization should be evaluated sentence wise. Each sentence should have an individual annotation score irrespective of other sentences.
- The human annotation score should be for the semantics of the summarized sentence and not for lexical(similar words) summarization.
- The summarization should be non-redundant and informative.
- The annotator should keep in mind about both the fluency and the adequacy of the summarized sentences and give one score for each sentence.
- The annotation scores are in the range of 1-5 and the details of each score is in the table below

Score	Meaning	Details
1	Very bad	The generated summary completely inaccurately/inappropriately summarizes the given text/paragraph
2	Bad	The generated summary, contains some relevant information from the text but is mostly incorrect / inaccurate
3	Borderline	The generated summary roughly summarizes the original text, but non-negligible pieces of information are missing or are incorrect
4	Good	The generated summary is a good summary of the original text with some minor (generally negligible) errors
5	Very good	The generated summary is a very appropriate summary of the given text/paragraph

2.2 Generates title as the summaries (Gigaword corpus)

Guidelines -

The following guidelines are for the human evaluation of the automatic abstractive text summarization task,

- Use the English long sentences/paragraphs in the first column as the source data and the short English titles in the second column as the target data in the given excel file.
- The titles are the machine summarized data which needs to be evaluated with human annotation scores.
- In case of models generating titles, do not to evaluate them as proper sentences (i.e., grammatically incorrect title is ok) .
- The summarization should be evaluated sentence wise. Each sentence should have an individual annotation score irrespective of other sentences.
- The human annotation score should be for the semantics of the summarized sentence and not for lexical(similar words) summarization.
- The summarization should be non-redundant and informative.
- The annotator should keep in mind about both the fluency and the adequacy of the summarized sentences and give one score for each sentence.
- The annotation scores are in the range of 1-5 and the details of each score is in the table below

Score	Meaning	Details
1	Very bad	The generated title is completely inappropriate for the given text/paragraph
2	Bad	The generated title corresponds somewhat to the original text but is mostly incorrect / inappropriate
3	Borderline	The generated title roughly corresponds to the original text, but there are non-negligible errors
4	Good	The generated title is a good title of the original text, with some minor (generally negligible) errors
5	Very good	The generated title is a very appropriate title for the given text/paragraph