**Metrics for patent numbers summarization**

We have evaluated all the below metrics on 1630 google patent number text summarization and mentioned all the details and their results.

We have measured all the metrics on two models named **hupd-t5-small** and **hupd-t5-base**.

**Note:**

1. Calculated all metrics score for abstract summary, claims summary and combined summary separately.
2. For any metrics evaluation, we need to the generated summaries against one or more reference summaries.
3. For abstract summary, we have taken normal input Abstract as the reference summary and output abstract as the generated summary.
4. For claims summary, we have taken normal input Claims as the reference summary and output claims as the generated summary.
5. For combined summaries, we have taken normal both the input Claims Summary and Abstract summary as the reference summary and output of combined summary as the generated summary.
6. **Rouge Score:**

The Rouge Score (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics designed to evaluate automatic summarization, especially for text. It's also frequently used in machine translation, text generation, and other natural language processing tasks where output can be compared to a reference text.

ROUGE-N: This measures the number of N-gram matches between the system-generated summary and the reference summary. N-grams are continuous sequences of N words in a sentence. For example, ROUGE-1 refers to the overlap of 1-gram (each word) between the system and reference summaries, while ROUGE-2 refers to the overlap of 2-grams (two consecutive words).

Rouge score returns two values: precision and recall. Precision is the percentage of the system-generated summary words that appear in a reference summary, while recall is the percentage of words in a reference summary that appear in the system-generated summary. Additionally, F1 Score, the harmonic mean of Precision and Recall, is also commonly used in reporting Rouge scores.

To calculate Rouge scores, you typically compare the generated summaries against one or more reference summaries. The calculation involves comparing n-grams (contiguous sequences of words) between the generated and reference summaries.

Here's an example table for calculating Rouge scores using the Rouge-1, Rouge-2, and Rouge-L metrics:

**Generated Summary** **Reference Summary 1** **Reference Summary 2**

This is a sample summary. This is a reference summary. Another reference summary.

The model performs well. The model performs nicely. The performance is good.

The scores range from 0 to 1, with higher scores indicating better overlap and similarity between the generated and reference summaries

1. **Bleu Score:**

The BLEU (Bilingual Evaluation Understudy) score is a widely-used metric for evaluating the quality of machine-generated text, especially in machine translation. The BLEU score quantifies the similarity between the machine-generated text and one or more reference texts.

It considers the matching n-grams between the generated text and the reference text. N-grams are contiguous sequences of 'n' words. It includes a brevity penalty to penalize short translations that may achieve higher precision by simply using fewer words. BLEU combines the scores for different n-grams (usually 1 to 4) into a single score by calculating the geometric mean of the modified precision counts. The final BLEU score ranges from 0 to 1; 1 indicates a perfect match with the reference, while 0 indicates no overlap in n-grams.

It’s important to note that BLEU is an automated metric and does not always perfectly reflect human evaluation.

The BLEU score is a metric used to evaluate the quality of machine translation outputs. It measures the similarity between a machine-generated translation and one or more reference translations, with scores ranging from 0 to 1.

1. **BERT Score:**

Based on the BERT model, a pre-trained language model that has been shown to be effective at a variety of natural language processing tasks.

BERTScore computes a similarity score between a generated text and a reference text by matching words in the two texts using contextual embeddings. BERTScore has been shown to correlate well with human judgment of text quality, and it has been used to evaluate a variety of text generation tasks, including machine translation, summarization, and question answering.

Some of the advantages of using BERTScore include that it is based on a pre-trained language model, which makes it less computationally expensive to compute than other evaluation metrics; it takes into account the semantic similarity between the two texts, in addition to the n-gram overlap; and it has been shown to correlate well with human judgment of text quality.

Some of the disadvantages of using BERTScore include that it is a relatively new metric, and it has not been as widely studied as other evaluation metrics; it can be sensitive to the choice of hyperparameters; and it is not always clear how to interpret the BERTScore score.

For accuracy, the score ranges from 0 to 1, where 1 represents perfect accuracy.

For F1 score, the score ranges from 0 to 1 as well, where 1 represents the best possible precision and recall trade-off.

Precision and recall scores also range from 0 to 1, with 1 indicating the highest precision or recall.

It's important to note that the scoring range can be influenced by factors such as class imbalance in the dataset or the specific threshold used for binary classification tasks.

1. **Flesch Reading Ease (FRE):**

The Flesch Reading Ease (FRE) score is a metric that assesses the readability of an English text. It tells you how easy or difficult a text is to read by examining the sentence length and word length. Here are the key points:

Calculation: It's calculated using the formula:

FRE = 206.835 - (1.015 \* ASL) - (84.6 \* ASW)

Where ASL is the average sentence length (i.e., the number of words divided by the number of sentences), and ASW is the average number of syllables per word (i.e., the number of syllables divided by the number of words).

**Scale:** The score typically ranges from 0 to 100. Higher scores indicate that the text is easier to read, while lower scores indicate that the text is more difficult to read.

**Interpretation:** Texts with a score above 90 are considered very easy and understandable by an average 5th grader. Scores between 60-70 are considered to be at the 8th and 9th-grade level. Academic and technical texts often score below 30 and are considered difficult to read.

**Usage:** It is widely used in content creation, education, and publishing to ensure that materials are appropriate for the intended audience's reading ability.

**Limitations:** The FRE score mainly focuses on sentence and word length and does not consider other factors such as content, coherence, or text structure which also impact readability. Also, it's specifically calibrated for English and may not be applicable to other languages.

The Flesch Reading Ease Score is typically interpreted as follows:

90-100: Very easy to read. Easily understood by an average 11-year-old student.

80-89: Easy to read. Conversational English for a wide audience.

70-79: Fairly easy to read. Simple language and plain English.

60-69: Plain English. Easily understood by 13- to 15-year-old students.

50-59: Fairly difficult to read. High school level.

30-49: Difficult to read. College level.

0-29: Very difficult to read. Graduate level.

These ranges provide a general guideline for interpreting the Flesch Reading Ease Score. Higher scores indicate easier readability, while lower scores indicate more complex or difficult-to-read texts. However, it's important to note that these ranges are not absolute and can vary depending on the specific context and target audience.

1. **Coleman-Liau Index (CLI):**

The Coleman-Liau Index (CLI) is a readability formula used to assess the difficulty level of a text. It provides an estimate of the grade level required to understand the text. The CLI score is calculated using the average number of characters per 100 words and the average number of sentences per 100 words.

The formula for CLI is: CLI = 0.0588 \* L - 0.296 \* S - 15.8, where L represents the average number of characters per 100 words, and S represents the average number of sentences per 100 words. It's important to note that the CLI score focuses on two factors: average sentence length and average word length. It doesn't take into account word difficulty or subject matter complexity. Therefore, it may not be suitable for assessing texts with specialized terminology or jargon.

The range of CLI scores varies from negative values to around 20 or higher. Lower scores indicate easier readability, while higher scores indicate more advanced and complex text.

CLI Score of 0-5: Very easy to read. Generally, corresponds to lower grade levels, such as elementary school.

CLI Score of 6-10: Easy to read. Corresponds to middle school or early high school level.

CLI Score of 11-15: Moderately difficult to read. Corresponds to high school or early college level.

CLI Score above 15: Difficult to read. Corresponds to advanced college or graduate level.

1. **Dale-Chall Readability (DCR):**

The Dale-Chall Readability (DCR) score is a metric used to assess the difficulty level of a text. It takes into account two main factors: word familiarity and sentence length. The DCR score calculates the percentage of difficult words in a text based on a list of 3,000 familiar words known to fourth-grade students. If a word is not on this list, it is considered difficult. Additionally, the DCR score considers the average sentence length in words.

The formula for calculating the DCR score is DCR = 0.1579 \* (PDW) + 0.0496 \* (ASL), where PDW represents the percentage of difficult words and ASL represents the average sentence length. The DCR score provides an estimated grade level equivalent, indicating the number of years of education required to comprehend the text. It is commonly used to assess the readability of educational materials and instructional texts, particularly in terms of vocabulary difficulty. However, it's important to note that the DCR score has limitations as it does not consider other factors such as sentence structure, content complexity, or contextual comprehension. Therefore, it should be used alongside other considerations and critical judgment when determining the suitability of a text for a specific audience or purpose.

The DCR score does not have widely established grade level ranges like some other readability metrics. However, here is a general guideline for interpreting the DCR score:

DCR Score below 5: Very easy to read. It corresponds to lower grade levels, such as elementary school.

DCR Score of 5-7: Easy to read. Corresponds to middle school or early high school level.

DCR Score of 8-9: Moderately difficult to read. Corresponds to high school or early college level.

DCR Score above 9: Difficult to read. Corresponds to advanced college or graduate level.

1. **N-gram:**

N-gram score metrics are used to assess the similarity or proximity between texts by examining sequences of words or characters called "n-grams." An n-gram is a contiguous sequence of n items, such as individual words or characters. N-gram score metrics calculate the overlap or similarity between the n-grams in two texts, providing a quantitative measure of their resemblance. These metrics are commonly employed in tasks like plagiarism detection, document comparison, and text categorization, as they capture the structural and lexical aspects of the texts being analyzed.

The size of the n-gram chosen depends on the specific application and the characteristics of the texts. Smaller n-grams, like unigrams or bigrams, focus on individual words or pairs of words, whereas larger n-grams, like trigrams or more, consider longer sequences to incorporate more contextual information. The choice of n-gram size strikes a balance between sensitivity to small variations (with smaller n-grams) and capturing meaningful patterns (with larger n-grams). N-gram score metrics offer a versatile approach to comparing and measuring textual similarity, enabling a range of applications in text analysis, information retrieval, and beyond.

The ranges of N-gram evaluation scores vary depending on the specific N-gram model and the corpus that it is trained on. However, in general, N-gram scores range from 0 to 1, with 1 being the best possible score. A score of 0 indicates that the N-gram model does not match any of the n-grams in the test corpus, while a score of 1 indicates that the N-gram model matches all of the n-grams in the test corpus. The following shows the ranges of N-gram evaluation scores for some common N-gram models:

N-gram Model Range

Unigram 0 to 0.9

Bigram 0 to 0.8

Trigram 0 to 0.7

4-gram 0 to 0.6

As you can see, the ranges of N-gram evaluation scores decrease as the order of the n-gram increases. This is because the higher-order n-grams are less common in natural language, so it is more difficult for an N-gram model to match them.

It is important to note that N-gram evaluation scores are not always a good indicator of the quality of an N-gram model. For example, an N-gram model that is trained on a very small corpus may have high N-gram scores, even if it is not a very good model. Therefore, it is important to use other measures of N-gram model quality, such as perplexity, in addition to N-gram evaluation scores.

1. **SummaC:**

The SUMMAC evaluation used various metrics to assess the performance of automatic text summarization systems. Here are the key evaluation metrics used:

Relevance Assessment: This metric focused on determining the accuracy of summaries in terms of relevance to the given topic. The evaluation measured the F-score, which is a combined measure of precision and recall, to quantify the agreement between the summaries and the full-text documents. A lower degradation in F-score indicated better performance.

Decision-making Time: This metric measured the time taken by analysts to make decisions based on the summaries. It compared the time required for decision-making using the full-text documents versus using the summaries. The reduction in decision-making time indicated the efficiency of the summarization system.

Informativeness: In the question-answering task, the informativeness of topic-related summaries was evaluated. Automatic methods were introduced to measure the informativeness of the summaries. The correlation between the scores given by these automatic methods and the informativeness scores provided by human judges was examined.

It's important to note that these metrics were specifically used in the context of the SUMMAC evaluation. Other evaluation efforts in the field of automatic text summarization may employ different or additional metrics depending on the specific objectives and requirements of the evaluation.

Ranges of SummaC evaluation scores:

0% - 50%: Low inconsistency

50% - 75%: Medium inconsistency

75% - 100%: High inconsistency

The SummaC evaluation score is a measure of the inconsistency between a summary and its source text. The score is calculated by comparing the summary to the source text and identifying any inconsistencies in the information. The scores are then grouped into three ranges, with a score of 0-50% indicating low inconsistency, a score of 50-75% indicating medium inconsistency, and a score of 75-100% indicating high inconsistency. It is important to note that the SummaC evaluation score is just one measure of inconsistency. Other factors, such as the length of the summary and the complexity of the source text, can also affect the score.