An Automated Quantitative Semantic Metric for Summarization Quality On Arbitrary Texts

A Preprint

Andrew D. Foland *
Department of Narrative Research
Odescape
Wellesley, MA 02482
andrew.foland@odescape.com

Abstract

This paper proposes Summarization Quality Using Embeddings (SQUE), a new quantitative objective metric for evaluation of summarization quality of arbitrary texts that relies on both the the closeness of semantic sentence embeddings and the summary length compression. This gives a measure of how well the recall-compression tradeoff is managed, the most important skill in summarization. Experiments demonstrate that SQUE effectively captures the token-length / semantic retention tradeoff of a summarizer and correlates to human perception of sumarization quality. It provides an automated alternative for assessing summarization quality without relying on time-consuming human-generated reference summaries. The proposed metric can be applied to various summarization tasks, offering a valuable automated tool for evaluating and improving summarization algorithms. summarization prompts, and synthetically-generated summaries.

Keywords Summarization · Semantic Embeddings · Metrics

1 Introduction

Quantitative methods for summarization capability evaluation today still often have drawbacks for the era of high-rate automated task completion via LLM's. Several methods have been devised for quantitative measures of summarization quality, as recently reviewed in Retkowski [2023] and Fabbri et al. [2021].

Skill at summarization is generally demonstrated by maximizing content retention for a given level of summary compression[2]. Indeed, as seen in figure 1, it becomes (unsurprisingly) increasingly difficult to retain all semantic content as the length of the summary is reduced.

Surprisingly, standard methods do not generally compare candidate summaries to the original text to be summarized. Most methods focus on precision or recall compared to a reference dataset of summaries *rather than to the text to be summarized itself*; yet these do not necessarily align to human skill in summarization; for instance recall score can easily be increased by writing a longer summary.

We therefore wish to incorporate an explicit comparison to the parent text, including a measure of token length compression, into our evaluation of summarization quality. Quantifying a good balance is desirable for automated summary generation, for instance by large language models (LLM's).

We aim to capture this skill as a measurable property of a summarization process, based only on the parent text and the candidate summary. To our knowledge SQUE is the first method to take into account the compression / recall tradeoff in a text/candidate context evaluating summarization quality. It is also, to our knowledge, the first method to measure that tradeoff motivated in a form with no free parameters or *ad hoc* terms.

*The author does not speak for Leidos, Inc.

[2]Famously, the short story commonly attributed to Hemingway: "For sale: baby shoes, never worn."

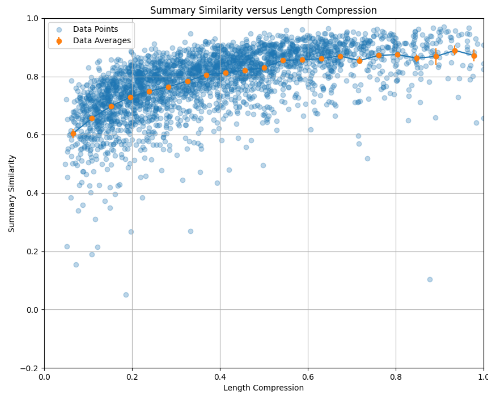

Figure 1: Summary embedding similarity to parent document similarity as a function of length compression. As the summarizer uses fewer tokens, semantic similarity is reduced. The quality of a summarizer can be measured by how quickly it falls off as a function of token compression.

We emphasize that the precise measure of semantic similarity, subject to numerical assumptions in form described below, can be changed without necessarily changing the nature of the SQUE metric. As semantic measures improve over time, SQUE is straightforwardly generalized to use them in place of the measures proposed in the present communication. We demonstrate this in the course of the present report by using an alternate embedding model.

2 Contributions

In this paper we contribute:

1. An objective, reference-free, no-free-parameters summarization metric comparing parent text and summary directly that is input-length independent and usable on arbitrary texts and summaries
2. A dataset of LLM-generated candidate summaries of 456 different texts, where the summaries vary in length and quality
3. A demonstration that the metric captures well the tradeoff of a summarizer between token length and semantic retention
4. Evidence for only weak embedder-dependence of the metric under particular numerical conditions as to the form of the similarity measure
5. Evidence for superior correlation of metric output values to human perception of summarization quality

3 Background

3.1 Previous Work

As discussed in Retkowski [2023]:

> Most commonly, summarization systems are evaluated on automated metrics. ROUGE Lin [2004] in particular has a long-standing history in the field and measures the lexical overlap between reference summaries and generated summaries. More recent metrics such as BertScore Zhang et al. [2019] and BARTScore Yuan et al. [2021], which are better at capturing semantic equivalence, are also becoming increasingly established.

One such is the ROUGE metric Lin [2004], which compares the generated summary to a reference summary and calculates the overlap between the two. The process of comparing a summary to reference summaries (and not the parent paragraph) using the ROUGE metric is described as "standardized" for summarization evaluation in Fabbri et al. [2021]. Another attempt is PRIMERA Xiao et al. [2022], which evaluates the content coverage and structural similarity of the generated summary to the reference summary. Many other methods are described in detail in Fabbri et al. [2021].

A method comparing the text and candidate summary in a similar spirit to this proposal is SUPERT **??**; however SUPERT does not take into account length, and relies on automated generation of extractive reference summaries from the text for comparison to the summary.

4 The Methodology

We wish to devise a metric for summarization quality. It is to be expected that there is always a tradeoff between faithfulness to the original text's full meaning, and the length of the summary. A good summary is one that retains as much fideliy as possible given its length. We formulate a metric that is based on a ratio of semantic retention to token compression.

4.1 Overview

Consider an idealized summarizer that compresses the token count T by a multiplicative compression factor k at each step and reduces its semantic similarity by a multiplicative degradation factor D at each step. High summarization quality corresponds to k small and D large, so the quantity $\frac{D}{k}$ (or any similar function monotonically increasing in D and monotonically decreasing in k) forms a measure of the summarization quality of this idealized summarizer. After N compressions, we have T_N tokens left:

$$T_N = T_0 k^N, \tag{1}$$

and semantic degradation D is given by

$$D_N = D_0 D^N = D^N \tag{2}$$

for $D_0 = 1$, which we are free to scale.

We would like a metric on the observables D^N and $\frac{T_N}{T_0}$ that is a function of $\frac{D}{k}$, when k and N are unknown. The metric M_{SQUE} that does this is:

$$M_{\text{SQUE}} = \frac{\ln D_N}{\ln \frac{T_N}{T_0}} = \left(\frac{\ln D}{\ln k}\right) \tag{3}$$

Note that is indeed increasing in D and k similarly to $\left(\frac{D}{k}\right)$. We keep it in the form $\frac{\ln D}{\ln k}$ as it is more suggestive than the reduction to $\log_k D$.

Note particularly that by taking this form, we do not need to know what N or k is–we measure a quality metric value directly rather than model parameters of a postulated summarization process. The postulated idealized summarizer model serves only to motivate and illuminate the form, which stands or falls on its own usefulness and correlation to perceived quality.

The form also means that the quality as measured is continuous–we are not restricted to integer N. Finally, we note that the metric goes smoothly to negative values–in the case where the summary is longer than the original text. Even in the perfect-semantic preservation case, one would expect this to have negative utility, as captured by the metric.

4.2 Measurement of Semantic Degradation

While counting tokens T_N is very straightforward, how are we to measure the semantic degradation D?

We propose to measure it by evaluating the similarity that derives from a semantic sentence embedder Felix Hill and Korhonen [2016], Wang et al. [2020], Sutskever et al. [2014], Muennighoff et al. [2022].

A sentence embedding is a dense vector representation of a sentence, which captures its semantic meaning in a continuous vector space. It is a form of word embedding, which has been extended to handle entire sentences. The concept of sentence embeddings has evolved from the need to move beyond word-level representations, as words can have different meanings based on context. The history of sentence embeddings can be traced back to the early 2000s, with foundational work on word embeddings like Word2Vec Mikolov et al. [2013] and GloVe Pennington et al. [2014], **?**. The development of recursive neural networks and, later, transformer-based models allowed for the creation of more sophisticated sentence embeddings, capable of capturing long-range dependencies and nuanced meanings. Notable techniques include contextualized embeddings from models like BERT Devlin et al. [2018] and RoBERTa Liu et al. [2019]. These advancements have significantly improved the performance of natural language processing tasks such as semantic textual similarity, paraphrase detection, and machine translation.

It is important to choose a sentence embedder and similarity metric with the following features:

- Semantically identical meanings should have a similarity measure of 1
- Semantically unrelated meanings should have a similarity measure of 0

A cosine-similarity measure of a sentence embedding vector will often have these features:

$$D \equiv \hat{v}_{summ} \cdot \hat{v}_{text}. \tag{4}$$

Note that should better measures than embeddings or cosine become available in the future, they can be fully incorporated into SQUE subject to the numerical conditions of limiting at 0 and 1, and functioning approximately multiplicatively.

4.3 Motivation for Cosine Similarity as Degradation Measure

Why should we expect degradation in cosine similarity to be multiplicative?

Degradation by D in cosine space corresponds to moving by some angle. And in particular, to moving by some "random" direction and length in embedding space.

If (by Ansatz) the lengths are constant at each step, but the angles are fixed (with rahdom direction, then the net random step is a length in some direction, along some distance $\ell_\perp$ and $\ell_\parallel$. Crucially, in high dimensions the random $\ell_\perp$ directions at different steps will be orthogonal to one another Krapivsky [2020]. So the accumulated orthogonal component grows like a random walk[3] of N orthogonal steps–aka like $\sqrt{N}$.

For small angles, the angle itself grows linearly with the orthogonal component. Hence the relevant function looks like $\cos \alpha\sqrt{N}$, where α is controlled by the amount of parallel and perpendicular components per step. This function looks strikingly like an exponential over most of its range–hence the Ansatz of multiplicative loss–as shown in figure 2. The range from $\cos\sqrt{x} = 1$ to $\cos\sqrt{x} = 0.2$ is the relevant one for summaries[4], as random text pairings can give rise to semantic similarities as large as 0.2.

To the extent that one can extend the discrete "summarization step" to a continuous "summarization operator" which can be applied at arbitrary strength, with 0-strength corresponding to the identity operator, the exponential function (multiplicative Ansatz) correponds to a linear and real operator generator.

4.4 Sentence Embeddings

Many sentence embedders use only a subset of the feature space Kennedy [2023], resulting in particularly the second condition (0 similarity for unrelated texts) being violated. We choose an embedder for which we can show both conditions hold.

We show that two commonly-used sentence embedders, both of which satisfy both conditions, yield broadly similar results when evaluating the summarization metric over a common dataset.

4.5 Summarization

Summarization is a common NLP task. Given the recent interest in large language models (LLM's) and automated text generation, interest is renewed in automated summzarization. Given this, it is cumbersome for human-dependent reference texts to form the backbone of summarization evaluation.

In this paper we target an objective metric of summarization based on semantic similarity. This metric does not depend on human judgment or creation of a good reference text.

[3] Note that the $\sqrt{N}$ behavior continues to hold in high dimensions.

[4] A true degradation operator would not take on negative values, suggesting that there may be a better candidate function of the embedding vectors than the cosine similarity with which to evaluate the semantic similarity. However as we will see it serves with the properties we desire over this range.

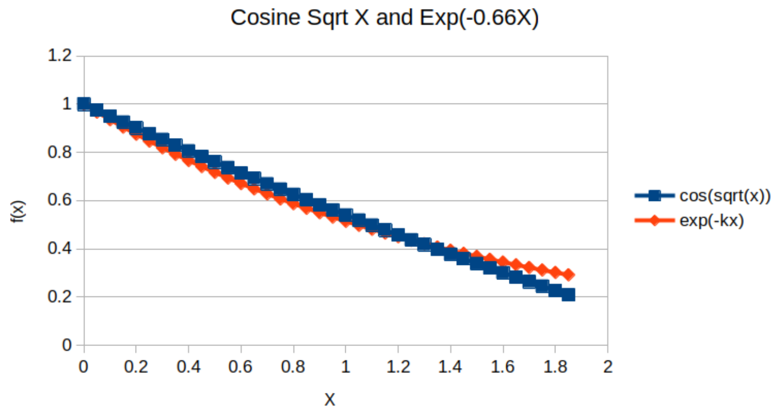

Figure 2: Demonstration of similarity of $\cos \sqrt{x}$ and $e^{-0.66x}$ over the range from $\cos \sqrt{x} = 1$ to $\cos \sqrt{x} = 0.2$. The coefficient value of 0.66 is found by a least-squares fit to the cosine over the range. The functional closeness motivates the choice of cosine similarity as a multiplicative semantic degradation measure.

4.6 Content and Length Variability

It would be unsurprising if summarization quality were content-dependent. For the purposes of this paper, we work to include a broad array of content and topics.

As sentences are quantized at the word level, it would not be unexpected that summarization quality measured by the metric would be dependent on the text length, particularly for low values of the text length. We investigate the dependence of the metric on this length.

4.6.1 Length Correlations

If the sentence embedding model encodes (either explicitly, or for instance as a distributed principal component) the input sequence length, then apparent semantic differences in summarization could in principle be a combination of perfect semantic agreement coupled with an encoded length difference. Similar questions have been investigated previously, e.g., in Adi et al. [2016]. We investigate a paraphrase-based test and other vector-component level tests to investigate this possibility.

5 Experimental Setup

We describe how we evaluate the utility of the proposed SQUE metric. As described previsouly, the metric stands or falls on its own usefulness and correlation to perceived quality. We therefore wish to demonstrate the following:

- Independence of embeddings from input length
- That the similarity metric we use is near 1 for semantically similar texts and near 0 for semantically unrelated texts
- That the summarization quality metric shows only weak dependence on the sentence embedding model (once the key 0-to-1 requirement is met)
- That the summarization quality metric is near 0 when the "summarization" of a text is unrelated to the text
- That the summarization quality metric metric forms a compact distribution away from 0 for summaries of different lengths known to derive from a parent unsummarized text
- That the summarization quality metric operates as expected not only on summaries, but also hierarchically on summaries of summaries
- That the metric correlates to known quantitative measures and qualitative perceptions of summarization quality.

Table 1: Topics and sources in the summarization dataset, taken from Khashabi et al. [2018].

	Sources	
Category	Sources	Approx. %
News	CNN, NYT, WSJ	15%
Articles	Wikipedia	10%
Articles on law and justice	Ide and Suderman 2006	10%
Articles on history and anthropology	Ide et al 2009	5%
Elementary School Science Textbooks	www.ck12.org	20%
9/11 Reports	Ide and Suderman, 2006	10%
Fiction	Project Gutenberg, Children's Stories, CMU Movie Summaries	30%

5.1 Datasets

5.1.1 Summarization Dataset

For investigations, we use the MultiRC2 (R2) dataset Khashabi et al. [2018]. The dataset consists of 456 paragraphs drawn from many different topics, styles, and sources. These are described in 1, as taken from Khashabi et al. [2018]. We convert each paragraph into a single clean text prior to embedding. Typical paragraph lengths are 200-500 words.

5.1.2 Paraphrase Dataset

For paraphrases, we desire a dataset with paragraphs that can usefully be paraphrased and are of typical paragraph size. Ideally they would be paragraphs no LLM has seen before. For this purpose we use Anonymous [2023] to seed a few-shot prompt for generation of random paragraphs. We obtain 6 seed paragraphs spanning different settings, styles, and points of view. We seed these as few-shot "random paragraphs" and the LLM (`LoneStriker_dolphin-2.5-mixtral-8x7b-6.0bpw-h6-exl2-2`) generates new random paragraphs. We harvest 490 paragraphs of typically 40-100 words randomly generated in this way.

We then proceed to paraphrase these 490 parapgraphs. We seed the 6 seed paragraphs, along with 8 manual paraphrases each, as a few-shot paraphrase exercise and the LLM (`LoneStriker_dolphin-2.5-mixtral-8x7b-6.0bpw-h6-exl2-2`) generates new paraphrase paragraphs as a continuation (*nb*, not as an instruction-following chat format.)

The fewshot continuation prompt is structured as an instructional text with examples, wherein the examples are of concise and verbose paraphrases of exemplar texts. The instructional heading of the prompt is

- **Paraphrase Fewshot prompt header**: "# What is Paraphrasing? Paraphrasing refers to the process of rephrasing a piece of text or information while retaining its original meaning. It is an essential tool used by writers, editors, and students to convey the same message in different ways, making it easier for readers to understand and remember the information. Paraphrasing is important because it helps to improve writing skills, avoid plagiarism, and enhance comprehension by presenting information in a more concise and clear manner. Additionally, paraphrasing can be used to tailor information to different audiences, making it more accessible and engaging for readers. Paraphrasing can be verbose and wordy, expanding the length of the passage, or concise and succinct, shortening the passage." *Followed by 6 example texts, each with 8 example paraphrases of varying lengths.*

5.2 Sentence Embedding

For sentence embedding, we use the `all-MiniLM-L6-v2` Wang et al. [2020] embedding model. This model is chosen for its widespread use and good properties. It can embed up to 512 tokens, typically about 400 words. The most important good property that is has, shown in figure 3, is that random pairs of text drawn from our text dataset show cosine similarity centered near 0. Many embeddings show much larger values of cosine similarity (often centered in the range 0.4-0.7) for random pairings of text samples. This may indicate Kennedy [2023] that such models are using only a subspace of the full embedding dimension.

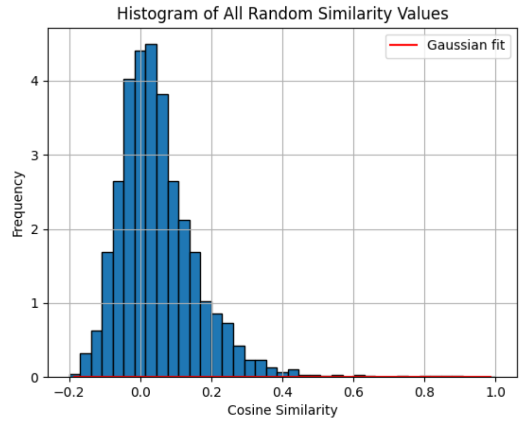

Figure 3: Distribution of semantic similarity for random pairings of paragraphs with the `all-MiniLM-L6-v2` sentence embedding.

5.3 Language Model and Prompts

We use a language model to generate summaries of our text dataset. The language model we use is `LoneStriker_Mixtral-8x7B-Instruct-v0.1-5.0bpw-h6-exl2`, a quantized version of a high-performing open-weights Mixture-of-Experts model. For each text, the model was hierarchically invoked and prompted to produce a summary, first of the original text, then of the summary of the original text, then of the summary of the summary. The model was prompted as follows:

- **Original text**: "Provide a concise 200-word summary of the key information in the following text. Provide the summary in the same voice and tense as the original. Do not add anything else. Provide only the requested 200-word summary:"
- **First summary**: "Provide a short, concise summary, of no more than 75 words, of the following text. Provide the summary in the same voice, tense, and view as the input text. Reduce the length significantly. Do not add anything else. Provide only the requested 75-word short summary:"
- **Second summary**: "Provide a one-sentence summary of the following text, retaining only the most important information. Reduce the length significantly. Provide the summary in the same voice and tense as the original text. Do not add anything else. Provide only a short, concise, one-sentence summary:"

5.4 Note on Token Count

When we give token counts in this paper, we give them based on the `tiktoken` token counter package from OpenAI OpenAI [2022]. This is neither the tokenizer used by our embedders (which are custom to each embedder) nor our LLM's (which use the llama tokenizer Touvron et al. [2023]). However, the various tokenizers tend to give similar counts within the precision of this paper ($\pm 20\%$) and are not expected to impact the conclusions, as token counts in the paper are used mainly to form compression ratios in which overall token count scale factors cancel.

5.5 Test Summarizations

To investigate summarization, we summarize the text hierarchically as described previously. Each summary is embedded with the sentence embedder. For each summary, the ratio of output summary tokens to input text tokens $\frac{T_{summ}}{T_{text}}$ is calculated for the denominator of the M_{SQUE} metric. The cosine similarity $\hat{v}_{text} \cdot \hat{v}_{summ}$ is evaluated as D for the numerator of the M_{SQUE} metric.

We show in figure 4 that the similarity of a text with its summary is typically near 1.

5.6 Null Summarizations

To demonstrate the the cosine similarity metric is near zero for unrelated texts, we randomly pair (exluding the correct pairings) texts with the summaries of other texts. The resulting distribution is shown in figure 3, peaking near 0.

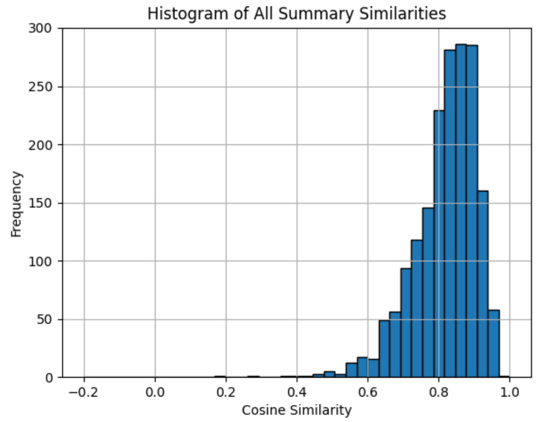

Figure 4: Distribution of cosine similarity for pairings of paragraphs with their summaries. The distribution is peaked near 1.0.

5.7 Paraphrase Tests

We investigate the length dependence and high-similarity behavior of the sentence embedding using a dataset of paraphrases. The paraphrases are generated by the `LoneStriker_dolphin-2.5-mixtral-8x7b-6.0bpw-h6-exl2-2` language model. The model was prompted few-shot with six paragraph examples of generating (for each paragraph) four concise and verbose paraphrases.

Four "concise" and four "verbose" paraphrases are generated for each input. Due to imperfections in the LLM, even the "concise" samples are often longer than the original text. We review all paraphrases manually, and all are acceptable examples of paraphrasing of the original.

5.7.1 Semantic Similarity Distribution

In figure 4 we showed that paraphrase similarity to input text is very high, typically larger than 0.9. In the high (368) dimensional space of the embedding, cosine similarity of 0.9 indicates very strong alignment.

In figure 5, we show a scatter plot of semantic similarity as a function of normalized paraphrase length. Over a range near unchanged paraphrase length, similarity is high and flat, falling off at more extreme differences. It is important to know whether the falloff is due truly to the small semantic differences that can exist between paraphrases (and which surely exist more when conditioned on large length difference) or whether there might be elements of the embedding that correlate directly to length of the embedded text. Such elements would confound measurement of an image quality metric combining semantic information with length information, over-incorporating the length information.

5.7.2 Evaluation of Text Length Correlations

We evaluate whether the embedding directly encodes the input length. Over all 4410 examples, we examine correlation of each of the 368 vector dimensions to both length and the normalized length (length divided by length of the original.) We believe normalized length is a better measure for this test to isolate the variations of semantically similar texts. The observed correlations are shown in figure 6.

For normalized length, we find the distribution of correlations is centered at zero with a standard deviation of the distribution of approximately 0.2. We find we cannot reject the null hypothesis that no component is correlated with length. We find no components with absolute value of correlation greater than 0.45. We conclude that single components with weak correlation are unlikely to induce significant shifts in the semantic cosine similarity due to token length variation alone.

Similarly, we see no pattern of strong correlations either with the raw length, or with any of the vector components of a principal components decomposition.

5.8 Code and Dataset Availability

All of the code and datasets used for this paper can be found at `https://github.com/afoland/sque`

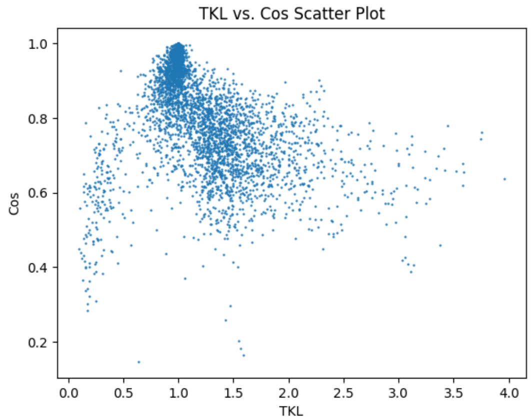

Figure 5: Distribution of cosine similarity for paragraphs and their paraphrases, as a function of normalized paraphrase length $\frac{\ell_{para}}{\ell_{text}}$.

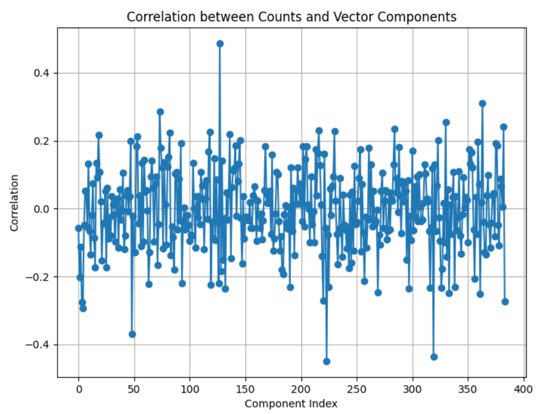

Figure 6: Observed correlations between 368 individial embedding vector components and the normalized length of the input. No correlation patterns strong enough to materially alter cosine similarity due to length alone are evident.

6 Results and Analysis

6.1 Performance Evaluation

First we show that summarization quality is very low and tightly clustered near 0 when the summary is not a summary of the parent paragraph, in Figure 7.

Next we show that the summarization quality is distinctively better and clustered significantly away from zero for true summarizations. The distribution of M_{SQUE} is shown in figure 8. The mean of the distribution 4.55±0.06 is separated by over two standard deviations of the distribution (and by many[5] standard errors on the mean).

The meaning of $M_{SQUE} = 4.55$ is that for each halving of a token length in summarization, the semantic degradation is $\ln 2/4.55$, or a multiplicative factor of 0.86. Note the good agreement of this value with the average cosine similarity (i.e., degradation) at 0.5, 0.25 ($0.86^2 = 0.74$), and 0.125 ($0.86^3 = 0.63$) in figure 1. This implies the metric captures well the tradeoff curve of the summarizer in length versus semantic retention.

[5]While we report the fit uncertainty value of 0.06 on the mean, we believe this may understate the uncertainty somewhat, as there may be correlations between, for instance, the M_{SQUE} evaluation of a text and summary, and between the summary and its own summary, due to being in similar parts of semantic embedding space. It is beyond the scope (and likely beyond the useful point of value) of the present report to precisely determine these correlations and a more accurate value of the fit uncertainty.

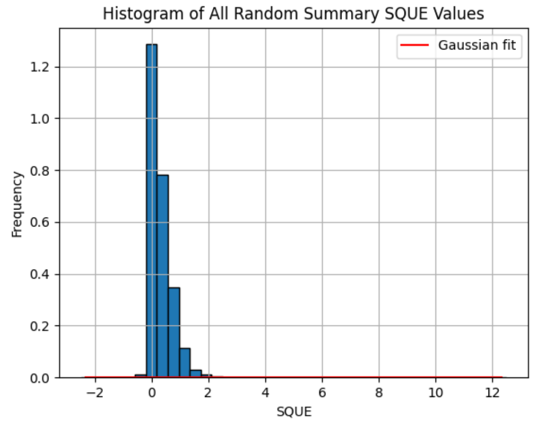

Figure 7: Distribution of M_{SQUE} quality for random pairings of paragraphs with summaries of other paragraphs. The mean value is 0.35, close to 0.

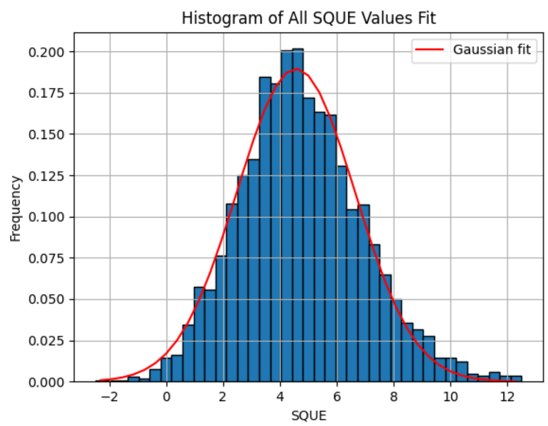

Figure 8: Distribution of M_{SQUE} quality for summaries, summaries of summaries, and summaries of summaries of summaries. Also shown is a gaussian fit to the distribution. The fitted mean value is 4.55±0.06, with a Gaussian σ of 2.08. It is well-separated from the values for random pairings of paragraphs and other summaries.

Finally, in figure 9 we show there is no meaningful length-dependent trend in quality. We find no evidence for trend on average, with an upper limit of no more than 10% variation per 500 tokens.

6.2 Hyperparameter Studies

6.2.1 Form of the SQUE Expression

As motivated, the expression for SQUE has no free parameters. However, as a test, we evaluate the case where the SQUE expression is modified by raising the denominator to a power:

$$M_{\mathrm{SQUE}} = \frac{\ln D_N}{-\ln |\frac{T_N}{T_0}|^p} \tag{5}$$

where p is a power in the range $[0, 2]$, and the absolute values are necessary for fractional p as the compression factor is nearly always negative.

In the limit $p \to 0$, we recover SQUE $\to \hat{v}_{summ} \cdot \hat{v}_{text}$. As $p \to 2$, the SQUE value becomes increasingly tolerant of semantic loss for highly-compressed summaries, leading to increasing overlap with random pairings of summaries.

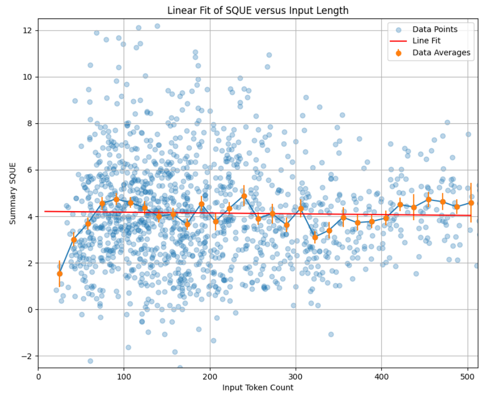

Figure 9: Distribution of M_{SQUE} against input text length (in tokens). In addition to the scatterplot, the average y value (M_{SQUE} value) is shown, together with its standard error on the mean. Also shown is a linear fit to the trend. The fitted slope is 0.0003±0.0005, or at most about 0.40 units of $MSQUE$ quality (about 10% variation) per 500 tokens

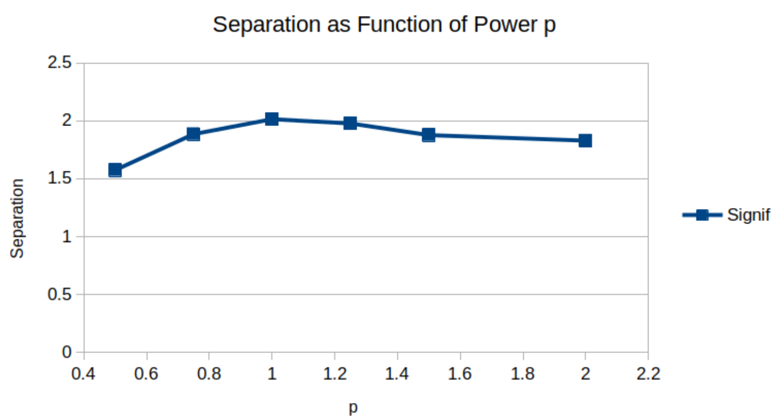

Figure 10: M_{SQUE} distribution separation (in standard deviations of the distribution) between random pairings of text, and true summaries, as a function of the power p applied to the denominator of SQUE. The separation peaks at 1.0, the originally-motivated value of p, and we conclude this is an empirically acceptable value.

We evaluate the separation of the random SQUE distribution and the summary SQUE distribution, $\frac{\mu_{summ}-\mu_{random}}{\sqrt{\sigma^2_{summ}+\sigma^2_{random}}}$. The resulting distribution, shown in figure 10, has a broad maximum, peaking at 1.0. We conclude that 1.0, as motivated, is empirically an acceptable point of operation.

6.2.2 Context-Dependence: Principal Components Analysis

To Do: PCA discussion

6.2.3 Using a Different Embedder

We repeat the analysis using the commonly used `all-mpnet-base-v2` Song et al. [2020]. The identical texts and summaries are used.

As shown in 11, this embedder satisfies the numerical requirements for similar and dissimilar texts. It also shows a similar-shaped falloff of semantic retention as a function of token compression. This embedder finds modestly higher SQUE values, that are generally correlated ($\rho = 0.55$) with the SQUE values found using `all-MiniLM-L6-V2`. The value of SQUE found is 5.8±0.09 with a standard deviation of 2.9; the distribution's separation from random texts (SQUE=0) is nearly identical (i.e. 2σ) to that of the default embedding model.

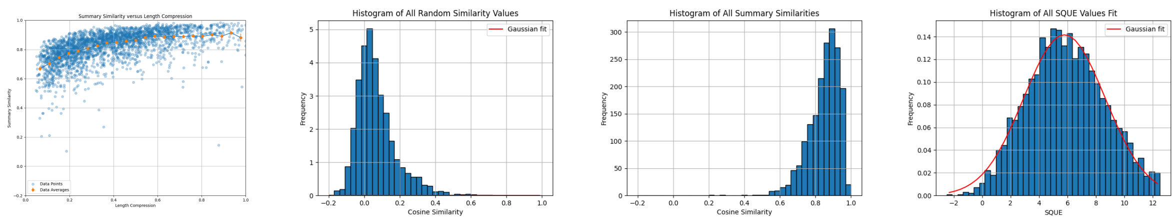

Figure 11: Analysis of the summary dataset using another embedder, `all-mpnet-base-v2`. From left to write: similarity as a function of token cmopression; similarity for random unrelated text pairings; similarity for summary pairs; distribution of observed M_{SQUE} values. The SQUE values are modestly higher for this embedder.

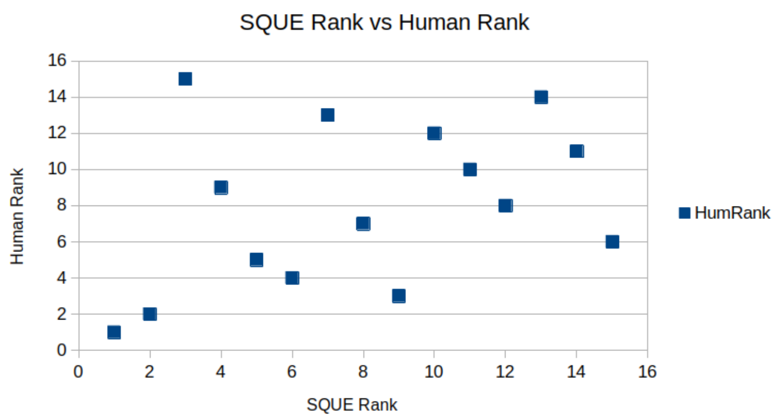

Figure 12: Human rank-ordering of test-summary pair summarization quality as a function of SQUE metric rank for the pair. The correlation is +0.36, demonstrating that SQUE correlates to human perception of summarization quality.

The higher SQUE value implies that more of the semantic content important to `all-mpnet-base-v2` is being retained than the semantic content important to `all-MiniLLM-L6-v2`. This is an example of the known differences among embedders in their sensitivity to different semantic aspects.

6.3 Comparison to Human Evaluation

As described previously, human evaluation of summary quality considers both semantic retention and compression. We wish to evaluate SQUE's correlation to human evaluation. We select 3 texts in each of 5 bins of compression([0.2,0.3],[0.3,0.4],[0.4,0.5],[0.5,0.6],[0.6,0.8]). The 3 texts selected are the texts nearest to the 10%, 50%, and 90% percentile in SQUE for that compression bin. The fifteen texts are presented blindly and in random order to the human evaluator for rank ordering in summarization quality.

The resulting rank correlations are shown in figure . Note that the top two summaries by SQUE were correctly ranked in order as the top two summaries by the human evaluator. The correlation coefficient of the rank ordering is 0.36. This is significantly larger than has been found for the widely-used ROUGE metric Liu and Liu [2010] (Table 1) and larger than most standard evaluation methods Gao et al. [2020] (Table 1). We conclude that SQUE with this embedding correlates with human perception of summarization quality.

6.4 Efficiency and Objectivity

The entire summarization dataset, consisting of 1824 embeddings of approximately 360,000 tokens is evaluated in less than 60 seconds on a Intel Xeon 28-core CPU. We conclude that evaluation of the metric is very fast, significantly faster than most LLM summarization processes.

The evaluation is objective insofar as it is fully automated and does not depend on any subjective evaluation or human input. The only inputs are the (arbitrary) text and its candidate summary.

7 Conclusion

7.1 Summary of Contributions

In this paper we have demonstrated

1. An objective, reference-free, parameter-free summarization metric that directly compares parent text and summary, independent of length and usable on arbitrary texts and summaries
2. That the metric captures well the tradeoff of a summarizer between token length and semantic retention
3. That the metric definitively distinguishes between summaries of texts, and random summary pairings of shorter lengths but unrelated meanings (the latter having quality near zero)
4. Indications of modest embedder-dependence of the metric (conditioned on numerical preconditions being met)
5. Superior correlation of metric output values to human perception of summarization quality

Harnessing sentence embedding techniques, SQUE moves beyond traditional lexical overlap methods and subjective reference summary requirements.

By comparing the semantic embeddings of both input documents and generated summaries, SQUE offers an efficient, objective, and content-coverage-focused evaluation method.

7.2 Potential Impact and Future Work

In the world of LLM-automated content generation and task completion, an automated, high-speed summarization quality metric is likely to play an important role in development.

Future work might include investigation of topic dependence and vulnerability/stability of the metric under a wider array of embeddings, particularly at long input lengths.

7.3 Real-World Applications

The proposed SQUE metric can be utilized to enable high-quality synthetic summary generation by filtering automatically generated summaries. Summarization algorithms can be employed to create a pool of candidate summaries for a given input document. The SQUE metric can be applied to each automatically generated summary, calculating the semantic embedding closeness scores. By setting a predefined threshold based on experimental analysis or domain knowledge, the system can filter out summaries with lower-quality scores, retaining only those that surpass the threshold.

The filtered high-quality summaries can then be utilized for various purposes, such as feeding back into the training process of summarization models, offering users a selection of high-quality summaries to choose from, or being used as reference summaries for further evaluation.

Periodic updates of the threshold and re-evaluation of generated summaries can help refine the summary generation process, leading to continuous improvement in summarization algorithm performance and generated summary quality. This filtering mechanism based on the SQUE metric ensures objective, efficient, content coverage-focused evaluation without the need for time-consuming human-generated reference summaries.

The metric can also be used when crafting prompts for summarization, as feedback to indicate what prompts are likely to be most successful in generating good summaries for a given task.

References

Fabian Retkowski. The Current State of Summarization. *arXiv e-prints*, art. arXiv:2305.04853, May 2023. doi:10.48550/arXiv.2305.04853.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, April 2021. ISSN 2307-387X. doi:10.1162/tacl_a_00373. URL `https://doi.org/10.1162/tacl_a_00373`.

Fabian Retkowski. The current state of summarization (arxiv 2305.04853), 2023.

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-1013`.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. September 2019. URL `https://openreview.net/forum?id=SkeHuCVFDr`.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. BARTScore: Evaluating Generated Text as Text Generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Paper.pdf`.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. PRIMERA: Pyramid-based Masked Sentence Pre-training for Multi-document Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.acl-long.360. URL `https://aclanthology.org/2022.acl-long.360`.

Kyunghyun Cho Felix Hill and Anna Korhonen. Learning distributed representations of sentences from unlabelled data. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *CoRR*, abs/2002.10957, 2020. URL `https://arxiv.org/abs/2002.10957`.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014. URL `http://arxiv.org/abs/1409.3215`.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022. doi:10.48550/ARXIV.2210.07316. URL `https://arxiv.org/abs/2210.07316`.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi:10.3115/v1/D14-1162. URL `https://aclanthology.org/D14-1162`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL `http://arxiv.org/abs/1907.11692`.

P. Krapivsky. https://mathoverflow.net/questions/320630/, 2020.

Curtis Kennedy. some-questions-about-text-embedding-ada-002-s-embedding, 2023.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *CoRR*, abs/1608.04207, 2016. URL `http://arxiv.org/abs/1608.04207`.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

Anonymous. Random paragraph generator: https://randomwordgenerator.com/paragraph.php, 2023.

OpenAI. https://github.com/openai/tiktoken, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models. 2023.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *CoRR*, abs/2004.09297, 2020. URL `https://arxiv.org/abs/2004.09297`.

Feifan Liu and Yang Liu. Exploring correlation between rouge and human evaluation on meeting summaries. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(1):187–196, 2010. doi:10.1109/TASL.2009.2025096.

Yang Gao, Wei Zhao, and Steffen Eger. SUPERT: towards new frontiers in unsupervised evaluation metrics for multi-document summarization. *CoRR*, abs/2005.03724, 2020. URL `https://arxiv.org/abs/2005.03724`.