The Likelihood Gain of a Language Model as a Metric for Text Summarization

Dana Levin
School of Computer Science
Reichman University
Email: dana.levin@post.runi.ac.il

Alon Kipnis
School of Computer Science
Reichman University
Email: alon.kipnis@runi.ac.il

***Abstract*—The gain in the log-likelihood (LLG) of a text under a language model (LM) when the text's summary is provided as a context to the LM, compared to no summary in the context, has been proposed as a reference-free index for the relevance of the summary to the text. We provide an information-theoretic interpretation of the LLG and an empirical analysis of the parts of speech affecting it most. We first show that the LLG describes the reduction in the binary codelength when the summary text is provided as side information to a lossless text compression system involving the LM and an entropy encoder. Consequently, under proper normalization, LLG is a form of the Normalized Compression Distance (NCD) and thus adheres to a universal information distance that is motivated by algorithmic information theory. Empirical results show that an NCD based on LLG is better correlated with human annotators than a gzip-based NCD. Additionally, we empirically show that LLG is affected almost exclusively by tokens associated with the text's content rather than tokens associated with its structure. Our findings support LLG as a natural and useful metric for evaluating text summarization methods.**

I. Introduction

A. Motivation

Text summarization can be viewed as a constrained information distillation problem in the following sense. The summary conveys a considerable amount of information about the original text but is of relatively short size and must be human-readable. The motivation for our study is the challenge of assessing the quality of a given summary or comparing several potential summaries. Better assessment may lead to the development of better summarization methods.

It is natural to capture the philosophical idea of text summarization via Shannon's information measure. Namely, for a random text T and a potential summarization procedure of T leading to the summary S, the relevance of the summary may be defined as the mutual information between the two:

$$I(T;S) = H(T) - H(T|S). \tag{1}$$

Assuming that the text is generated by a language model (LM), the sample equivalent of the entropy is the negative log-likelihood under this model (also known as the logloss, negative log-perplexity [1], or cross-entropy with the token-indicator distribution). Consequently, a natural index for the relevance of a summary s to a text $t = t^n = (t_1, \ldots, t_n)$ is:

$$\mathrm{LLG}(t,s) := \ell(t) - \ell(t|s), \tag{2}$$

where we denoted

$$\ell(t|s) := \ell(t|s;\mathrm{P_{LM}}) := -\sum_{i=0}^{n} \log\left(\mathrm{P_{LM}}(t_i|\mathrm{ctx}(s,t^{i-1})\right),$$

and $\ell(t) = \ell(t|\emptyset)$. Here, the LM $\mathrm{P_{LM}}$ provides a probability distribution over the i-th token t_i given the context. This context is procured by the "context policy" function ctx that receives as inputs the previous tokens $t^{i-1} = (t_1, \ldots, t_{i-1})$ and potentially the summary text s. We defer the specification of the function ctx to the parts of this paper involving numerical evaluations. Ideologically, LLG quantifies the contribution of a summary s to the ability of the LM to predict the text t. To the best of our knowledge, a version of (2) for evaluating text summarization was first proposed in [2] under the name "information difference" (ID). The term LLG appears to be better suited in the context of information theory, hence we use it instead of ID.

In this work, we study properties of (2) in the context of lossless text compression that supports its usage as a measure of text summarization.

B. Background on Text Summarization Evaluation

Summarization procedures are roughly divided into extractive, in which the summary is limited to be a subset of the text, and abstractive, in which there is no such limitation. Automatic methods for summary evaluation include approaches based on lexical matching such as ROUGE [3], sentence embedding such as BERTscore [4], and question answering [5].

Another category of evaluation method is founded on the principles of Shannon's next token guessing game [6]. A guesser is tasked with predicting a document's content, one token at a time, based on other parts of the document and potentially the summary text [7]. The quality of a summary is measured by the improved prediction compared to a setup when the summary is not provided to the guesser. Methods falling under this category include the bidirectional fill-in-the-blank method of [8] and the LLG of (2) proposed in [2]. More specifically, the authors of [2] provided several measures for text summarization based on different LLG normalizations and strategies to form the context in (2). They empirically showed that these measures are generally better correlated with human evaluation scores than competing approaches. We also note that Shannon's guessing game principle was recently used in

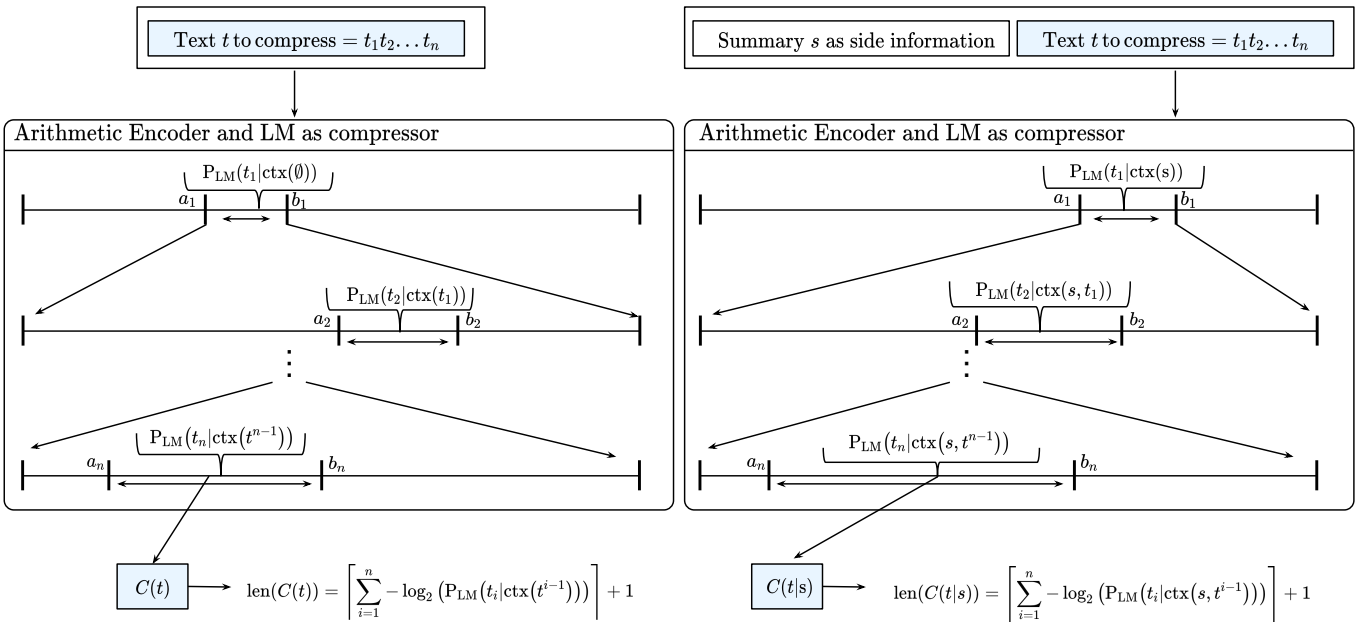

Fig. 1. Lossless text compression (encoder only) using a language model (LM) followed by an arithmetic encoder. Left: The sequence of tokens $t^n = (t_1, \dots, t_n)$ is encoded using an arithmetic encoder with token distributions given by the LM $\mathrm{P_{LM}}$. Right: the summary text s is provided as side information that modifies the LM next-token distribution through the context policy ctx. The reduction in codelength is due to a potentially improved prediction of the tokens by the LM. The logloss gain $\mathrm{LLG}(t, s)$ is within one bit of the difference between the binary codelength without and with side information.

a multimodal setup [9], and is also related to ranking words in the context of keyword extraction [10].

C. Contributions

We provide an information-theoretical interpretation and empirical justification for the LLG in the context of text summarization. We first show that LLG can be viewed as the reduction in the binary codelength when compressing a text t using a lossless compression procedure when the summary text s is provided as side information. This compression procedure involves a LM (that may be pre-trained or trained on the fly [11]) followed by an arithmetic encoder as illustrated in Figure 1. This architecture is known to attain state-of-the-art compression on large texts [12]. As a result of this characterization of LLG, we conclude that a normalized version of it can be seen as a normalized compression distance (NCD) which is a well-studied concept in information complexity and retrieval [13], [14]. Consequently, a comparison to the popular `Gzip`-based NCD is called upon. We conduct such a comparison and show that our implementation of LLG attains a better correlation with human evaluation scores. Finally, we perform Part of Speech (POS) analysis that shows that LLG is more influenced by tokens representing content rather than structure. This finding provides a linguistic justification for LLG as a similarity measure for summary evaluation.

In conclusion, LLG adheres to a universal distance measure that naturally considers the text's content but is relatively unaffected by its structure.

D. Organization

In Section II we discuss lossless compression. In Section III we show that LLG is a form of NCD. In Section IV we study LLG via content and structure token analysis. Concluding remarks are in Section V. The code for all empirical results is available at https://github.com/LevinDana/LLG.

II. Lossless compression with side information

In this section, we describe a lossless text compression procedure with side information. We show that LLG is the reduction in binary codelength the summary provides when used as side information for compressing the text via this procedure.

The encoding process, described in Figure 1, is as follows. Let $t = t^n = (t_1, \dots, t_n)$ be the text and s the summary text. Assume that the LM provides a probability distribution over a finite dictionary of tokens given the context produced by ctx. We start by obtaining the distribution $\mathrm{P_{LM}}(\cdot|\mathrm{ctx}(s))$ and finding the interval in that distribution which corresponds to the true word probability $\mathrm{P_{LM}}(t_1|\mathrm{ctx}(s))$. We will denote the interval $[a_1, b_1)$. Next we partition $[a_1, b_1)$ by the distribution $\mathrm{P_{LM}}(\cdot|\mathrm{ctx}(s, t_1))$ and find the interval $[a_2, b_2)$ corresponding to t_2. We continue in this fashion until we reach t_n, partition and find its corresponding interval by $\mathrm{P_{LM}}(\cdot|\mathrm{ctx}(s, t^{n-1}))$ and $\mathrm{P_{LM}}(t_n|\mathrm{ctx}(s, t^{n-1}))$.

To get the encoded representation of t, we use, for simplicity, the Shannon-Fano-Elias method of taking the first $\lceil \ell(t|\mathrm{ctx}(s)) \rceil + 1$ bits in the binary representation of the midpoint of the final interval $[a_n, b_n) \subset [0, 1]$, excluding the leading 0 [15]. This representation attains within two bits of

the shortest prefix-free binary code which falls within the final interval $[a_n, b_n)$ (In principle, prefix-freeness is not needed here because our process encodes the entire text in one pass. In practice, however, it is useful since we may concatenate several blocks as [11], [16].). We denote the final binary representation by $C(t|s)$. Note that $C(t|s)$ depends both on the LM and the way ctx processes s and previous tokens to form a context. The original t can be recovered without error given its encoded representation $C(t|s)$, the LM $\mathrm{P_{LM}}$ with the context policy function ctx, and s; see the decoder of the similar encoding procedures described in [11], [17]–[19]. We denote by $C(t)$ the binary representation obtained under a similar procedure but without s in the context as illustrated in the left-hand side of Figure 1.

Our first result states that the reduction in binary codelength due to the summary in the context is up to one bit from the LLG, regardless of the length of the text.

Theorem 1: Denote by $C(t)$, $C(t|s)$ the encoded binary representation of t and t given s, respectively, obtained through arithmetic coding and Shannon-Fano-Elias binary codeword representation. Then

$$|\mathrm{len}\,(C(t)) - \mathrm{len}\,(C(t|s)) - \mathrm{LLG}(t,s)| \le 1 \tag{3}$$

Proof: By construction, we have

$$\begin{aligned}\mathrm{len}\,(C(t)) &= \lceil \ell(t) \rceil + 1\\ \mathrm{len}\,(C(t|s)) &= \lceil \ell(t|s) \rceil + 1.\end{aligned}$$

By the definition of $\mathrm{LLG}(t,s)$ in (2),

$$\begin{aligned}\mathrm{LLG}(t,s) - 1 &\le \lceil \ell(t) \rceil - \lceil \ell(t|s) \rceil\\ &\le \mathrm{LLG}(t,s) + 1.\end{aligned}$$

This implies (3). ■

III. Logloss Gain as a Compression Distance

The Normalized Compression Distance (NCD) is a universal similarity measure between two data objects x and y, regardless of their domain, obtained by analyzing the objects' compressed forms using a compressor Z [14]. Denoting by $Z(xy)$ the length (in bits) of jointly compressing x and y under z, the NCD is defined as

$$\mathrm{NCD}_Z(x,y) = \frac{Z(xy) - \min\{Z(x), Z(y)\}}{\max\{Z(x), Z(y)\}} \tag{4}$$

As a typical example, Z is a universal compression algorithm like `gzip` [20] and $Z(xy)$ is the length of the compression of x concatenated with y [21]. NCD has been used in various applications such as clustering [22] and across multiple domains such as text [23], music [22], and genomics [21]. We refer to [24] for discussions on the detailed implementation of NCD using `gzip` and other off-the-shelf compressors.

Consider

$$\overline{\mathrm{LLG}}(t,s) := \frac{\ell(t) - \ell(t|s)}{\ell(t)}, \tag{5}$$

which we denote as the Normalized LLG. The following theorem says that $1 - \overline{\mathrm{LLG}}(t,s)$ is a form of NCD under the compressor described in Section II.

Theorem 2: Consider the text compressor C described in Section II. For a pair of texts t and s with $\mathrm{len}\,(C(s)) \le \mathrm{len}\,(C(t))$, let $C(s,t)$ be the codeword obtained by first compressing s using $\mathrm{len}\,(C(s))$ bits and then compressing t with s as side information. We have

$$\overline{\mathrm{LLG}}(t,s) = 1 - \mathrm{NCD}_C(s,t).$$

Proof: When jointly compressing t and s using the process described earlier and in Figure 1, we get $\mathrm{len}\,(C(s,t)) = \mathrm{len}\,(C(s)) + \mathrm{len}\,(C(t|s))$. It follows that

$$\begin{aligned}\mathrm{NCD}_C(s,t) &= \frac{\mathrm{len}\,(C(s,t)) - \min\{\mathrm{len}\,(C(s)), \mathrm{len}\,(C(t))\}}{\max\{\mathrm{len}\,(C(s)), \mathrm{len}\,(C(t))\}}\\ &= \frac{\mathrm{len}\,(C(s,t)) - \mathrm{len}\,(C(s))}{\mathrm{len}\,(C(t))}\\ &= \frac{\mathrm{len}\,(C(s)) + \mathrm{len}\,(C(t|s)) - \mathrm{len}\,(C(s))}{\mathrm{len}\,(C(t))}\\ &= \frac{\mathrm{len}\,(C(t|s))}{\mathrm{len}\,(C(t))} = 1 - \overline{\mathrm{LLG}}(t|s).\end{aligned}$$

■

Remark 1: The work of [2] studied $\mathrm{LLG}(t,s)$ divided by $\ell(t) - \ell(t|t)$. This normalization coincides with (5) when $\ell(t|t) = 0$. Namely, when the LM utilizes the context to attain perfect next token prediction. In practice, $\ell(t|t)$ may be much smaller than $\ell(t)$ but typically far from zero.

A. Empirical Study

The empirical results in this paper are obtained using GPT2 [25] as the LM and with a context policy ctx that processes one sentence at a time. Namely, the context of the i-th token in the j-th sentence is the text summary s concatenated with the $i-1$ tokens in that sentence. This implementation of the LLG follows that of [2].

We use the following datasets:

- `SummEval` [26]. This dataset contains 1,700 summaries of CNN/Daily Mail articles. The dataset is composed of 100 text articles. Each article is paired with 17 summaries generated from 17 different models, 4 extractive and 13 abstractive. The summaries are scored by 3 human evaluators over 4 categories: Relevance, Coherence, Consistency, and Fluency. We focus here only on Relevance as it is ideologically closest to information among these categories. Relevance also enjoys a relatively high agreement among the 3 evaluators with κ agreement coefficient of 0.71 [27, Ch. 11]; when evaluated in a one-against-many fashion, the average correlation coefficient between the evaluators is $r = 0.6$ for Pearson and $\tau = 0.46$ for Kendall's-τ.
- `SFF` (Summarize from feedback) [28]. This dataset contains 6312 summaries of CNN/Daily Mail articles. Each article is scored by a human evaluator over 3 categories:

Correlation method	Dataset	
	SummEval	SFF
KENDALL τ	0.55 (0.011)	0.56 (0.007)
PEARSON	0.77 (0.012)	0.82 (0.007)

Table I. Pairwise correlations between $1-\mathrm{NCD}_{\mathtt{gzip}}(s,t)$ and the normalized logloss gain $\overline{\mathrm{LLG}}(t,s)$ of (5) over two datasets containing many (text, summary) pairs; bootstrapped standard errors are in brackets.

Measure	Correlation method	Dataset	
		summEval	SFF
$\overline{\mathrm{LLG}}$	KENDALL τ	**0.25** (0.016)	**0.29** (0.009)
	PEARSON	**0.37** (0.020)	**0.41** (0.010)
$\mathrm{NCD}_{\mathtt{gzip}}$	KENDALL τ	0.18 (0.017)	0.26 (0.009)
	PEARSON	0.27 (0.020)	0.34 (0.010)

Table II. Comparing measures of text summary relevance by their agreement with human evaluations. Pairwise correlation (Kendall τ, Pearson) of the normalized LLG $\overline{\mathrm{LLG}}$, and one minus the normalized compression distance based on gzip ($1-\mathrm{NCD}_{\mathtt{gzip}}$), with human evaluation scores in two datasets (`summEval` [26], `SFF` [28]). Highest values are in **bold**; bootstrapped standard errors are in brackets.

In this sentence content words are in red and structure words are in blue.

Fig. 2. Distinguishing between content and structure words based on parts-of-speech.

Measure	Correlation method	Dataset	
		summEval	SFF
LLG	KENDALL τ	0.24 (0.017)	0.20 (0.010)
	PEARSON	0.34 (0.021)	0.27 (0.012)
$\mathrm{LLG}_{\mathsf{cont}}$	KENDALL τ	0.22 (0.017)	0.21 (0.010)
	PEARSON	0.32 (0.020)	0.29 (0.012)
$\mathrm{LLG}_{\mathsf{stru}}$	KENDALL τ	0.18 (0.017)	0.12 (0.009)
	PEARSON	0.28 (0.022)	0.17 (0.013)

Table III. Comparing measures of text summary relevance by their agreement with human evaluations. Pairwise correlation (Kenadal τ, Pearson) of negative log-likelihood gain (LLG), Content-only LLG, and Structure-only LLG with human evaluation scores in two datasets (`summEval` [26], `SFF` [28]). The correlation between human evaluations and LLG is almost identical to the correlation with Content-only LLG, indicating that LLG is typically not affected by words associated with the text's structure.

Coverage, Coherence, and Accuracy. We focus on Coverage as it ideologically appears to be the closest to information among these categories.

In Table I we report on both Kendall's τ and Pearson's correlation coefficients between $\overline{\mathrm{LLG}}(t,s)$ and $1-\mathrm{NCD}_{\mathtt{gzip}}(s,t)$ over the two datasets above. The relatively high correlations between these measures indicate that the redundancies in the text introduced by the summary as measured by both methods are proportional to each other.

In Table II, we compare summarization scores obtained via $\overline{\mathrm{LLG}}(t,s)$ and $\mathrm{NCD}_{\mathtt{gzip}}(s,t)$ to human evaluation scores. Our results show that $\overline{\mathrm{LLG}}$ generally outperforms $\mathrm{NCD}_{\mathtt{gzip}}$. This appears to indicate that the benefit of semantic language understanding, unique to an LM-based compressor, gives $\overline{\mathrm{LLG}}(t,s)$ an advantage over general-purpose Lempel-Ziv-based compression.

IV. Content and Structure Information Analysis

In some linguistics and information retrieval studies, it is useful to distinguish between words associated with the text's *structure* or style, typically denoted as "function" words, to words associated with its *content* [29]–[31]. Figure 2 demonstrates such separation. Although summarization is conceptually associated with the text's content, the LLG of (2) considers all tokens of the text regardless of their association to structure or content. In what follows, we analyze the contribution of each token type to the LLG and empirically show that structure tokens affect very little on LLG. Namely, LLG tends to ignore information conveyed by the text's structure which is a desirable property of a summarization evaluation metric.

We label each token as "Content" or "Structure" based on their part-of-speech (POS). We include in the Content category nouns, proper nouns, verbs (excluding auxiliary verbs), adjectives, adverbs, and numbers; see the example in Figure 2. The resulting separation is a useful approximation to the distinction between content and function words that does not require training and is computationally efficient [30], [32]. The average proportion of content (respectively, structure) tokens across all texts in our study is about 0.55 (respectively, 0.45) with a standard deviation across documents smaller than 0.1. Namely, these proportion varies only slightly from text to text.

For a given text t, let $\mathrm{I}_{\mathsf{cont}} = \mathrm{I}_{\mathsf{cont}}(t)$ be the set of indices of content tokens, and let $\mathrm{I}_{\mathsf{stru}} = \mathrm{I}_{\mathsf{stru}}(t)$ be the set of indices of structure tokens (i.e., not content tokens). Define

$$L_{\mathrm{cont}}(t) := := \sum_{i\in \mathrm{I}_{\mathsf{cont}}} -\log(\mathrm{P}_{\mathrm{LM}}(t_i|\mathrm{ctx}(t^{i-1})))$$
$$L_{\mathrm{cont}}(t|s) := := \sum_{i\in \mathrm{I}_{\mathsf{cont}}} -\log(\mathrm{P}_{\mathrm{LM}}(t_i|\mathrm{ctx}(s,t^{i-1})))$$

and similarly define $L_{\mathrm{stru}}(t)$ and $L^{\mathrm{stru}}(t|s)$. Note that in either case, the context of a token is based on *all* tokens in t occurring before that token. We define the LLG variants:

$$\mathrm{LLG}_{\mathsf{cont}} := L_{\mathrm{cont}}(t) - L_{\mathrm{cont}}(t|s),$$
$$\mathrm{LLG}_{\mathsf{stru}} := L_{\mathrm{stru}}(t) - L_{\mathrm{stru}}(t|s),$$

For all text t and s we have

$$\mathrm{LLG}(t,s) = \mathrm{LLG}_{\mathsf{cont}}(t,s) + \mathrm{LLG}_{\mathsf{stru}}(t,s).$$

Figure 3 depicts the histograms of LLG, $\mathrm{LLG}_{\mathsf{cont}}$, and $\mathrm{LLG}_{\mathsf{stru}}$ over all text-summary pairs per dataset. It follows from this figure that LLG values are similar to $\mathrm{LLG}_{\mathsf{cont}}$ and are significantly less affected by $\mathrm{LLG}_{\mathsf{stru}}$. We further compare all LLG variants as summarization metrics by checking their correlation with human evaluation scores. Table III shows that LLG correlates with human evaluations very similarly to $\mathrm{LLG}_{\mathsf{cont}}$ and both outperform $\mathrm{LLG}_{\mathsf{stru}}$. This behavior suggests that LLG naturally removes the impact of information related to the text's style since, in contrast to content information, tokens associated with the style are about equally predictable whether or not side information is present.

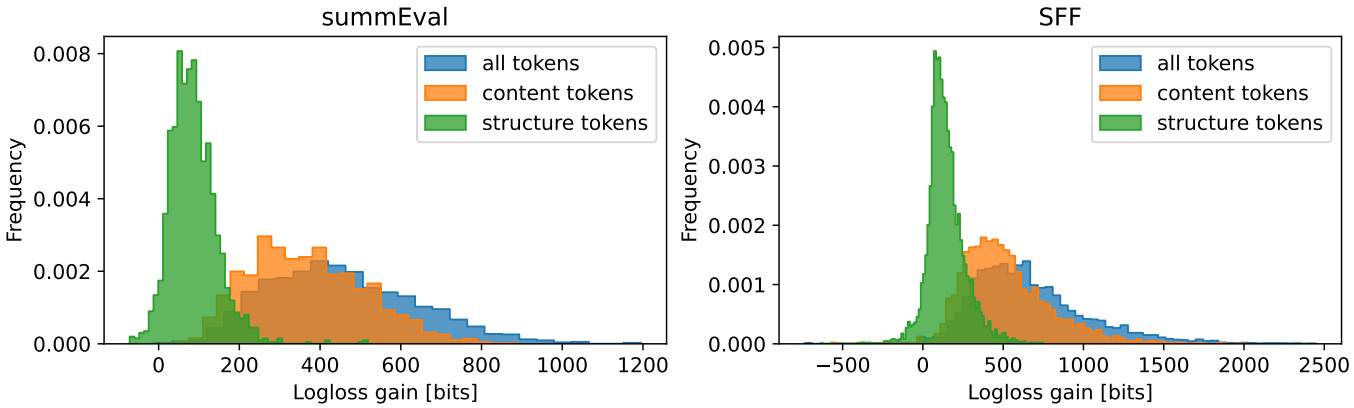

Fig. 3. Histograms of LLG, LLG_{cont} and LLG_{stru} over individual texts and their summaries in the `sumEval` dataset [26] (left) and `SFF` dataset [28] (right). LLG is the sum of LLG_{cont} and LLG_{stru}. The proportion of content/structure tokens across all texts is about $(0.55, 0.45)$ with a standard deviation across documents smaller than 0.1. These histograms imply that LLG is almost entirely affected by content tokens rather than tokens associated with the text's style.

We note that although small, the correlation of LLG_{stru} with human evaluations is significantly larger than zero, implying that our set of structure tokens provides some non-zero logloss gain on average. This can also be seen by the non-zero mean of LLG_{stru} values in Figure 3. Decompositions of LLG to components that do not correlate with human annotators may improve correlations of the complementary component and hence are desirable. We leave the search for such decompositions to future work.

V. Conclusions

We provided an information-theoretic interpretation of the gain in the LLG in the context of text summarization. Additionally, we showed that LLG naturally focuses on the content of the text rather than its structure, providing a form of linguistic justification for using it for scoring the relevance of summarization. These results suggest that LLG is a natural and useful index of similarity for evaluating and designing text summarization methods.

References

[1] D. Jurafsky and J. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, ser. Prentice Hall series in artificial intelligence. Pearson Prentice Hall, 2009.

[2] N. Egan, O. Vasilyev, and J. Bohannon, "Play the Shannon game with language models: A human-free approach to summary evaluation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10 599–10 607.

[3] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81.

[4] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," 2020.

[5] T. Scialom, S. Lamprier, B. Piwowarski, and J. Staiano, "Answers unite! unsupervised metrics for reinforced summarization models," in *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019, pp. 3237–3247.

[6] C. E. Shannon, "Prediction and entropy of printed english," *Bell system technical journal*, vol. 30, no. 1, pp. 50–64, 1951.

[7] E. Hovy and C.-Y. Lin, "Automated text summarization and the Summarist system," in *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*. Baltimore, Maryland, USA: Association for Computational Linguistics, Oct. 1998, pp. 197–214. [Online]. Available: https://aclanthology.org/X98-1026

[8] O. Vasilyev, V. Dharnidharka, and J. Bohannon, "Fill in the blanc: Human-free quality estimation of document summaries," 2020.

[9] V. Zouhar, S. Bhattacharya, and O. Bojar, "Multimodal shannon game with images," *arXiv preprint arXiv:2303.11192*, 2023.

[10] A. Tsvetkov and A. Kipnis, "Entropyrank: Unsupervised keyphrase extraction via side-information optimization for language model-based text compression," in *ICML 2023 Workshop Neural Compression: From Information Theory to Applications*, 2023.

[11] G. Izacard, A. Joulin, and E. Grave, "Lossless data compression with transformer," 2019. [Online]. Available: https://bellard.org/nncp/nncp.pdf

[12] M. Mahoney, "Large text compression benchmark," 2023. [Online]. Available: http://www.mattmahoney.net/dc/text.html

[13] M. Li, X. Chen, X. Li, B. Ma, and P. M. Vitányi, "The similarity metric," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3250–3264, 2004.

[14] M. Li, P. Vitányi *et al.*, *An introduction to Kolmogorov complexity and its applications*. Springer, 2008, vol. 3.

[15] T. Cover and J. A. Thomas, "Elements of information theory," 2006.

[16] F. Bellard, "Nncp v2: Lossless data compression with transformer," 2021.

[17] M. Goyal, K. Tatwawadi, S. Chandak, and I. Ochoa, "Dzip: Improved general-purpose lossless compression based on novel neural network modeling," in *2021 Data Compression Conference (DCC)*. IEEE, 2021, pp. 153–162.

[18] Y. Mao, Y. Cui, T.-W. Kuo, and C. J. Xue, "A fast transformer-based general-purpose lossless compressor," *arXiv preprint arXiv:2203.16114*, 2022.

[19] G. Deletang, A. Ruoss, P.-A. Duquenne, E. Catt, T. Genewein, C. Mattern, J. Grau-Moya, L. K. Wenliang, M. Aitchison, L. Orseau *et al.*, "Language modeling is compression," in *The Twelfth International Conference on Learning Representations*, 2023.

[20] P. Deutsch, "Rfc1951: Deflate compressed data format specification version 1.3," 1996.

[21] R. L. Cilibrasi and P. M. Vitányi, "Fast phylogeny of sars-cov-2 by compression," *Entropy*, vol. 24, no. 4, p. 439, 2022.

[22] R. Cilibrasi and P. Vitanyi, "Clustering by compression," 2004.

[23] A. J. Pinho, D. Pratas, and P. J. S. G. Ferreira, "Authorship attribution using relative compression," in *2016 Data Compression Conference (DCC)*, 2016, pp. 329–338.

[24] M. Cebrián, M. Alfonseca, and A. Ortega, "The normalized compression

distance is resistant to noise," *IEEE Transactions on Information Theory*, vol. 53, no. 5, pp. 1895–1900, 2007.

[25] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:160025533

[26] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "Summeval: Re-evaluating summarization evaluation," 2021.

[27] Y. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete multivariate analysis: Theory and practice*. Springer Science & Business Media, 2007.

[28] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano, "Learning to summarize from human feedback," 2022.

[29] F. Mosteller and D. L. Wallace, *Applied Bayesian and classical inference: the case of the Federalist papers*. Springer Science & Business Media, 2012.

[30] P. A. Howarth, *Phraseology in English academic writing: Some implications for language learning and dictionary making*. Walter de Gruyter, 2013, vol. 75.

[31] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, and D. Woodard, "Surveying stylometry techniques and applications," *ACM Computing Surveys (CSuR)*, vol. 50, no. 6, pp. 1–36, 2017.

[32] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 human language technology conference of the north american chapter of the association for computational linguistics*, 2003, pp. 252–259.