Play the Shannon Game With Language Models: A Human-Free Approach to Summary Evaluation

Nicholas Egan, Oleg Vasilyev, John Bohannon

Primer AI
{negan, oleg, bohannon}@primer.ai

Abstract

The goal of a summary is to concisely state the most important information in a document. With this principle in mind, we introduce new reference-free summary evaluation metrics that use a pretrained language model to estimate the information content shared between a document and its summary. These metrics are a modern take on the Shannon Game, a method for summary quality scoring proposed decades ago, where we replace human annotators with language models. We also view these metrics as an extension of BLANC, a recently proposed approach to summary quality measurement based on the performance of a language model with and without the help of a summary. Using transformer based language models, we empirically verify that our metrics achieve state-of-the-art correlation with human judgement of the summary quality dimensions of both coherence and relevance, as well as competitive correlation with human judgement of consistency and fluency.

1 Introduction

With the ever-expanding development of new summarization algorithms in the NLP community, metrics that reliably measure summary quality are more important than ever. And yet, the most popular method for summary quality estimation remains the ROUGE (Lin 2004) family of metrics, which require human written reference summaries for comparison and measure summary quality through simple token overlap, ignoring the syntax and semantics governing the way humans use language.

The goal of a summary is to concisely state the most important information conveyed by a document. Examining summarization through this lens, one should be able to determine summary quality by measuring how much information from the document is represented in the summary. Put another way, when comparing alternative summaries of similar length, the information we gain from reading the original document should be minimal given the best summary.

The idea of measuring this difference in information content was proposed as the Shannon Game by Hovy and Lin (1998): they assign 3 humans the task of guessing a document letter by letter, where the first human is allowed to look at the document, the second human is allowed to look at a summary of the document, and the third human is given nothing at all. By measuring how many tries it takes the second human to guess the document compared to the other humans, you can evaluate how much information about the document is communicated in the summary, and therefore measure how good the summary is.

Contributions This paper proposes a new summarization evaluation metric, the *Shannon Score*, that performs the Shannon Game with a language model such as GPT-2 (Radford et al. 2019). By using a language model to autoregressively generate a document both with and without a summary as a prompt, we measure the information provided by the summary. One can view this method as a more theoretically driven extension to the recently proposed BLANC metric (Vasilyev, Dharnidharka, and Bohannon 2020), which measures the accuracy of unmasking document tokens with and without a summary. In addition to the Shannon Score, we also propose a variant we call *Information Difference*.

To understand the empirical performance of this method as a summary evaluation technique, we performed experiments to correlate our metrics against human judgement. We found that our metrics perform strongly on the SummEval benchmark (Fabbri et al. 2021), achieving state-of-the-art correlation with human judgement of summary coherence and relevance, and competitive correlation with human judgement of summary consistency and fluency.

2 Methods

2.1 Computing Information

Language models are probability distributions over documents, giving us $p(\mathcal{D})$ for some document $\mathcal{D}$. Autoregressive language models do this by predicting next token probabilities given prior tokens, modeling

$$p(x_t|x_1, \ldots, x_{t-1})$$

where our input document is broken into tokens $\{x_1, \ldots, x_n\}$. The Shannon information content, or surprisal, of event E with probability $p(E)$ of happening is defined as $I(E) = -\log p(E)$, so we can compute the information of a document according to our language model as

$$\begin{aligned} I(\mathcal{D}) = &- \log p(x_1) - \log p(x_2|x_1) - \ldots \\ &- \log p(x_n|x_1, x_2, \ldots x_{n-1}) \end{aligned}$$

2.2 Conditional Information

Suppose we had a conditional language model $p(\mathcal{D}|\mathcal{S})$ that gives us a probability distribution of documents that could correspond to a given summary $\mathcal{S}$. Using this conditional language model, we could compute the conditional information content $I(\mathcal{D}|\mathcal{S})$ as the amount of information we gain from the document $\mathcal{D}$ if we are already given the information of summary $\mathcal{S}$.

If $\mathcal{S}$ is a satisfactory summary of $\mathcal{D}$, then $I(\mathcal{D}|\mathcal{S}) < I(\mathcal{D})$, as documents that have little to do with the summary should be much less likely than documents that are relevant to the summary after conditioning the language model. If the summary fluently describes people, ideas, or relationships that appear in the document, then that should decrease the information one learns from subsequently reading the document.

Thus we can define an *Information Difference* metric of summary quality as:

$$ID(\mathcal{D}, \mathcal{S}) = I(\mathcal{D}) - I(\mathcal{D}|\mathcal{S})$$

The Information Difference tells us the change in document information between using the summary and not using the summary, and it is equivalent to the log likelihood ratio between the document and the document given the summary. While it is unbounded, it should be positive unless a summary does such a bad job that it makes the document more confusing to read.

Considering the fact that the summary that best preserves the information of a document is the document itself, we can view $I(\mathcal{D}|\mathcal{D})$ as a lower bound on $I(\mathcal{D}|\mathcal{S})$. Since this idea of having a third evaluator who has the document itself as help is inspired by the Shannon Game, we can compute the *Shannon Score* metric as:

$$s(\mathcal{D}, \mathcal{S}) = \frac{I(\mathcal{D}) - I(\mathcal{D}|\mathcal{S})}{I(\mathcal{D}) - I(\mathcal{D}|\mathcal{D})}$$

The Shannon Score gives us the ratio between how helpful the summary was and how helpful the document itself was. While this formula in theory is unbounded, it usually should be in the range 0 to 1, unless the summary makes the document more confusing or somehow explains the document better than the document itself.

2.3 Approximating Conditional Information

To the extent of our knowledge, there is no easy way to exactly condition a pretrained language model such as GPT-2 on a summary, even though there has been work on conditioning language models on fixed control codes (Keskar et al. 2019), bags of words, or discriminators (Dathathri et al. 2020). We also have a strong motivation not to train such a model because we want our method to be universal and robust, while summarization datasets are much smaller and more restricted in domain than the massive datasets that modern language models require.

We approximate $p(\mathcal{D}|\mathcal{S})$ by computing the probability that $\mathcal{D}$ is generated when we provide $\mathcal{S}$ as a prompt to a language model. We intuitively justify this idea by the fact that in real-world documents the most important information is often summarized at the top as an introduction, and then described in more detail in body paragraphs. This setup resembles the BLANC-help metric (Vasilyev, Dharnidharka, and Bohannon 2020), which measures language model token unmasking accuracy for a document when a summary is prepended. An alternative setup would be to finetune a language model on the summary which was also explored by Vasilyev, Dharnidharka, and Bohannon (2020), but we don't explore that method in this paper. We use the GPT-2 small language model (Radford et al. 2019) for our experiments, but investigate the use of other language models in section 5.1.

An issue we run into when computing information with GPT-2 is that the model can only be given a maximum of 1024 tokens, making many documents too large to fit in at once. To get around this, we approximate document information with an independence assumption between sentences in the document, meaning that only the preceding tokens within a sentence are provided when generating the next token in the sentence. In section 5.2, we investigate the effects of prompting the language model with additional upstream sentences of context.

3 Understanding Our Metrics

3.1 Information Visualization

A toy illustration of our methodology is shown in Figure 1. We picked a document excerpt in the CNN/DailyMail (Hermann et al. 2015) dataset and paired it with two abstractive summaries we wrote. While both of these summaries are grammatically correct and mostly consist of words from the document, one of the summaries is of high quality and the other is of low quality. The figure shows the information content of each token in the document as estimated by GPT-2 in 4 scenarios: $I(\mathcal{D})$ (the document on its own), $I(\mathcal{D}|\mathcal{D})$ (the document given the document), $I(\mathcal{D}|\mathcal{S})$ (the document given a summary) for the high quality summary, and $I(\mathcal{D}|\mathcal{S})$ for the low quality summary. A darker background color denotes higher information according to the model.

As you can see, the model gained less information from words like "gray" and "Varvara" after seeing those words in the high quality summary. We can also see that words like "Pacific" and "journey," which do not appear in the high quality summary, became more likely to appear in the document due to their association with concepts in the summary. The low quality summary may have helped the model predict words like "CNN," but it is unhelpful for words like "mammal" and "website" that are confusingly used in the summary. Very little information was gained from reading a document that was already read, except for the first token or two for each sentence. This is an artifact of our autoregressive language modeling setup, so measuring $I(\mathcal{D}|\mathcal{D})$ is useful for normalizing our Shannon Scores.

We used a truncated document and toy summaries here to demonstrate the Shannon Score in a concise way, but we included visualizations of real, full-length documents and summaries from the SummEval dataset in the appendix.

$I(\mathcal{D}) = 580$

(CNN) A North Pacific gray whale has earned a spot in the record books after completing the longest migration of a mammal ever recorded . The whale , named Var v ara , sw am nearly 14 , 000 miles (22 , 500 kilometers), according to a release from Oregon State University , whose scientists helped conduct the whale - tracking study . Var v ara , which is Russian for " Bar bara ," left her primary feeding ground off Russia 's S akh alin Island to cross the Pacific Ocean and down the West Coast of the United States to B aja , Mexico . Var v ara 's journey surpassed a record listed on the Guinness Worlds Records website . It said the previous record was set by a hump back whale that sw am a mere 10 , 190 - mile round trip

$I(\mathcal{D}|\mathcal{D}) = 52$

(CNN) A North Pacific gray whale has earned a spot in the record books after completing the longest migration of a mammal ever recorded . The whale , named Var v ara , sw am nearly 14 , 000 miles (22 , 500 kilometers), according to a release from Oregon State University , whose scientists helped conduct the whale - tracking study . Var v ara , which is Russian for " Bar bara ," left her primary feeding ground off Russia 's S akh alin Island to cross the Pacific Ocean and down the West Coast of the United States to B aja , Mexico . Var v ara 's journey surpassed a record listed on the Guinness Worlds Records website . It said the previous record was set by a hump back whale that sw am a mere 10 , 190 - mile round trip

---

$I(\mathcal{D}|\mathcal{S}) = 482$ for this high quality summary:

Varvara the gray whale traveled from Russia to Mexico, a swim of record breaking length.

(CNN) A North Pacific gray whale has earned a spot in the record books after completing the longest migration of a mammal ever recorded . The whale , named Var v ara , sw am nearly 14 , 000 miles (22 , 500 kilometers), according to a release from Oregon State University , whose scientists helped conduct the whale - tracking study . Var v ara , which is Russian for " Bar bara ," left her primary feeding ground off Russia 's S akh alin Island to cross the Pacific Ocean and down the West Coast of the United States to B aja , Mexico . Var v ara 's journey surpassed a record listed on the Guinness Worlds Records website . It said the previous record was set by a hump back whale that sw am a mere 10 , 190 - mile round trip

$I(\mathcal{D}|\mathcal{S}) = 540$ for this low quality summary:

The round humpback has told CNN mammals that Baja was a previous Pacific website for "Guinness."

(CNN) A North Pacific gray whale has earned a spot in the record books after completing the longest migration of a mammal ever recorded . The whale , named Var v ara , sw am nearly 14 , 000 miles (22 , 500 kilometers), according to a release from Oregon State University , whose scientists helped conduct the whale - tracking study . Var v ara , which is Russian for " Bar bara ," left her primary feeding ground off Russia 's S akh alin Island to cross the Pacific Ocean and down the West Coast of the United States to B aja , Mexico . Var v ara 's journey surpassed a record listed on the Guinness Worlds Records website . It said the previous record was set by a hump back whale that sw am a mere 10 , 190 - mile round trip

Figure 1: A comparison of token-wise information content within a document as estimated by GPT-2 in 4 scenarios: the document on its own, the document given the document, the document given a high quality summary, and the document given a low quality summary. Tokens with a darker background color have more information.

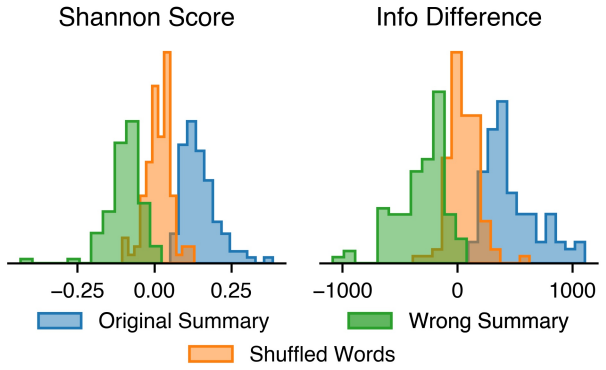

Figure 2: Distributions of Shannon Score and Information Difference on 100 summaries from the CNN/DailyMail dataset. Three different summaries are used: the original human written reference summary (in blue), the original summary with words scrambled (in orange), and a reference summary for a different document in the dataset (in green).

3.2 Baseline Validation

As a simple validation of our information-based metrics, we sampled 100 documents with their corresponding reference summaries from the CNN/DailyMail dataset (Hermann et al. 2015), and created two "bad" summaries per document: a version of the reference summary with all the words randomly shuffled, and a reference summary for a different document in the dataset.

Figure 2 shows the distributions of the Shannon Score and Information Difference for these three summaries. As expected, the original summaries have the highest scores, followed by shuffled summaries and wrong summaries. It is good to see that there is full separation between original summaries and wrong summaries for both metrics. The fact that the original summaries and shuffled summaries are almost completely separated demonstrates the importance of syntax to our metrics, a quality that metrics like the Jensen-Shannon divergence (Louis and Nenkova 2009) and ROUGE-1 (Lin 2004) lack.

We also verified that there are no documents for which the shuffled summary or wrong summary score better than the original summary for either of the metrics. Despite the fact that the Shannon Score has no lower bound, we can see that it doesn't go far below zero for even the most unreasonable of summaries. And despite the fact that the Shannon Score has no upper bound, even high quality human refer-

ence summaries are unable to achieve a score above 0.4.

4 Evaluation of Our Metrics

4.1 SummEval

Metric	Coher.	Consi.	Fluen.	Relev.
Shannonˆ	**0.4118**	**0.6324**	**0.5240**	**0.6029**
Info Diffˆ	**0.4706**	**0.6324**	**0.5683**	**0.6618**
rouge-1	0.2500	0.5294	**0.5240**	0.4118
rouge-2	0.1618	0.5882	0.4797	0.2941
rouge-3	0.2206	**0.7059**	0.5092	0.3529
rouge-4	0.3088	0.5882	**0.5535**	0.4118
rouge-L	0.0735	0.1471	0.2583	0.2353
rouge-su*	0.1912	0.2941	0.4354	0.3235
rouge-w	0.0000	0.3971	0.3764	0.1618
rouge-we-1	0.2647	0.4559	0.5092	**0.4265**
rouge-we-2	-0.0147	0.5000	0.3026	0.1176
rouge-we-3	0.0294	0.3676	0.3026	0.1912
S^3-pyr	-0.0294	0.5147	0.3173	0.1324
S^3-resp	-0.0147	0.5000	0.3321	0.1471
BertScore-p	0.0588	-0.1912	0.0074	0.1618
BertScore-r	0.1471	**0.6618**	0.4945	0.3088
BertScore-f	0.2059	0.0441	0.2435	**0.4265**
MoverScore	0.1912	-0.0294	0.2583	0.2941
SMS	0.1618	0.5588	0.3616	0.2353
SummaQAˆ	0.1176	0.6029	0.4059	0.2206
BLANCˆ	0.0735	0.5588	0.3616	0.2647
SuPERTˆ	0.1029	0.5882	0.4207	0.2353
BLEU	0.1176	0.0735	0.3321	0.2206
CHRF	**0.3971**	0.5294	0.4649	**0.5882**
CIDEr	0.1176	-0.1912	-0.0221	0.1912
METEOR	0.2353	**0.6324**	**0.6126**	**0.4265**
Lengthˆ	-0.0294	0.4265	0.2583	0.1618
Novel 1ˆ	0.1471	-0.2206	-0.1402	0.1029
Novel 2ˆ	0.0294	-0.5441	-0.3469	-0.1029
Novel 3ˆ	0.0294	-0.5735	-0.3469	-0.1324
Repeat 1ˆ	**-0.3824**	0.1029	-0.0664	**-0.3676**
Repeat 2ˆ	**-0.3824**	-0.0147	-0.2435	-0.4559
Repeat 3ˆ	-0.2206	0.1471	-0.0221	-0.2647
Coverageˆ	-0.1324	0.3529	0.1550	-0.0294
Compressˆ	0.1176	-0.4265	-0.2288	-0.0147
Densityˆ	0.1618	**0.6471**	0.3911	0.2941

Table 1: Kendall tau-b system-level correlation between expert annotations of coherence, consistency, fluency, and relevance and various automated metrics, adapted from Fabbri et al. (2021). ˆ denotes reference-free metrics. The five highest correlations per column are in bold, with ties for consistency and relevance. Coefficients with a magnitude above 0.36 are significant at the $\alpha = 0.05$ level.

The SummEval (Fabbri et al. 2021) benchmark was established as a comprehensive evaluation tool for summary evaluation metrics. It consists of 100 English-language documents from the CNN/DailyMail dataset, each paired with system summaries from 17 different summarization systems: 3 extractive models, 13 abstractive models, and a lead-3 baseline. All models were published in 2017 or later. Each of these 1700 system summaries were scored by a panel of 3 experts in the field of summarization on the qualities of coherence (the collective quality of all sentences), consistency (the factual alignment between the summary and document), fluency (the quality of individual sentences), and relevance (selection of important content from the source). The experts achieved an inter-annotator agreement kappa coefficient of 0.7127.

Fabbri et al. (2021) scored each summary using these evaluation metrics: ROUGE (Lin 2004), ROUGE-WE (Ng and Abrecht 2015), S^3 (Peyrard, Botschen, and Gurevych 2017), BertScore (Zhang et al. 2020), MoverScore (Zhao et al. 2019), Sentence Mover's Similarity (SMS) (Clark, Celikyilmaz, and Smith 2019), SummaQA (Scialom et al. 2019), BLANC (Vasilyev, Dharnidharka, and Bohannon 2020), SUPERT (Gao, Zhao, and Eger 2020), BLEU (Papineni et al. 2002), CHRF (Popović 2015), METEOR (Lavie and Agarwal 2007), and CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015). They also measure the Grusky, Naaman, and Artzi (2018) statistics of summary length, extractive fragment coverage (coverage), compression ratio, average length of extractive fragments (density), proportion of n-grams in summary that aren't in the document (novel n), and n-grams repeated in summary (repeat n).

Table 1 shows the correlation between expert annotations and the automated evaluation metrics. Following Fabbri et al. (2021), we use Kendall tau-b system-level correlation for comparison. Our metrics of Shannon Score and Information Difference are the only metrics to be in the top 5 for every category of summary quality. Additionally, our metrics achieve state-of-the-art performance for the qualities of coherence and relevance.

4.2 Coverage

The coverage score (Lin and Hovy 2003) is a human evaluation method that measures a system summary's recall of semantic units that appeared in a reference summary, weighed by how well the system summary was able to capture each semantic unit as judged by the human labeler. The 2001 and 2002 Document Understanding Conferences (DUC) provide datasets of English-language system and reference summaries for news documents with human coverage labels, on both single-document and multi-document levels.

Table 2 shows the correlation of various metrics to these coverage scores for the single-document summaries. System-level Spearman correlation is used following Louis and Nenkova (2013). The reference-free metrics perform similarly, except for Jensen-Shannon Divergence (Louis and Nenkova 2009) which performs particularly well on DUC 2001 and Info Diff which performs particularly poorly on DUC 2002. The metrics using references benefit from the bias that the coverage itself was measured with respect to the reference summary, so as expected, they have higher correlations with the coverage than the reference-free metrics for this dataset. A fair comparison would involve a coverage measured with respect to the document itself. One can also see that most metrics perform better on DUC 2002 than

Metric	DUC 2001	DUC 2002
Shannon Score	0.2909	0.5714
Info Diff	0.3000	0.4835
Jensen-Shannon	0.4455	0.5440
BLANC-help	0.2727	0.5769
ROUGE-1	0.9636	0.9066
ROUGE-2	0.8273	0.9121
ROUGE-L	0.7455	0.9176
ROUGE-Lsum	0.9455	0.9066
BERTScore-P	0.4636	0.5989
BERTScore-R	0.8545	0.9451
BERTScore-F1	0.6091	0.7308

Table 2: System-level Spearman correlation of various summary quality metrics with human-judged coverage scores on the DUC 2001 and 2002 single-document summary datasets. The last seven metrics make use of reference summaries, while the first four metrics have to rely only on the original document itself. DUC 2001 coefficients above 0.60 and DUC 2002 coefficients above 0.55 are significant at the $\alpha = 0.05$ level.

DUC 2001: this was also observed by Sun and Nenkova (2019), who suggested that this can be explained by the fact that DUC 2001 systems are more similar to each other and worse than DUC 2002 systems on average.

4.3 Metric Biases

To understand the biases of our metrics, we measured the correlation between our metrics and the SummEval statistics describing summaries described in section 4.1 across the 1700 SummEval summaries. For comparison, we also correlated the expert summary quality judgements with the statistics. These correlations are shown in table 3.

Both of our metrics have significant positive correlation with summary length, which is expected since longer summaries can contain more information. Our metrics have bias against more abstractive summaries (based on novel n-gram, coverage, and density), but we are generally less biased against abstractive summaries than humans judging consistency are: we suspect this is because abstractive summaries are more likely to hallucinate factual errors. The Shannon Score is biased against highly compressed summaries, which is not shared by Information Difference.

5 Metric Variations

5.1 Choice of Language Model or Model Size

In the previous sections, we used GPT-2 small as our language model of choice when computing the Shannon Score and Information Difference. To understand how well our method generalizes to other language models, we computed the Shannon Score and Information Difference metrics using the three other GPT-2 sizes (medium, large, and extra-large), and three other language models with autoregressive pretraining objectives: GPT (Radford et al. 2018), XLNet (Yang et al. 2019), and Transformer-XL (Dai et al. 2019).

Model	Coher.	Consi.	Fluen.	Relev.
Shannon Score				
GPT-2 S	**0.4118**	**0.6324**	**0.5240**	**0.6029**
GPT-2 M	0.3529	**0.6618**	0.4945	0.5441
GPT-2 L	0.3676	**0.6471**	0.5092	0.5588
GPT-2 XL	0.3824	**0.6324**	0.4945	0.5735
GPT	0.0294	0.5147	0.3469	0.1912
XLNet	**0.4265**	0.5882	0.4945	**0.6471**
TransfoXL	0.3529	0.5441	0.4502	0.5441
Information Difference				
GPT-2 S	**0.4706**	**0.6324**	**0.5683**	**0.6618**
GPT-2 M	0.3971	**0.6765**	0.5092	0.5882
GPT-2 L	0.3824	**0.6324**	0.4945	0.5735
GPT-2 XL	0.3971	**0.6471**	0.5092	0.5882
GPT	0.0441	0.5294	0.3616	0.2059
XLNet	0.4559	0.5882	0.5240	**0.6765**
TransfoXL	0.3529	0.5441	0.4502	0.5441

Table 4: Kendall tau-b system-level correlations between expert annotations of coherence, consistency, fluency, and relevance and our Shannon Score and Information Difference metrics with the choice of different language models on the SummEval dataset. Scores at least as high as GPT-2 S are bold. Coefficients above 0.36 are significant at the $\alpha = 0.05$ level.

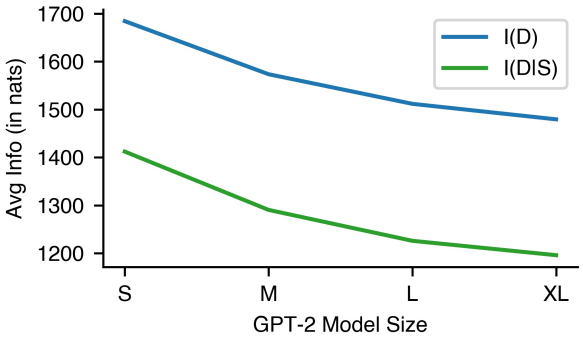

Figure 3: The average document information and document information given summary as estimated by different sizes of GPT-2 for the SummEval dataset.

Table 4 shows the system-level Kendall tau-b correlation between our metrics and the SummEval quality judgements from section 4.1 for each language model. The language models perform quite similarly overall, suggesting that the choice of language model is not overly important when using the Shannon Score or Information Difference. The exception is the low correlation of GPT, particularly on the coherence and relevance qualities: we suspect this is because GPT was trained on the BooksCorpus dataset (Zhu et al. 2015), which is less diverse than the datasets used for the other language models.

It is also interesting to see that bigger GPT-2 models do not necessarily perform better. Figure 3 shows the relation-

Metric	Info Diff	Shannon Score	Coherence	Consistency	Fluency	Relevance
Length	0.5425	0.4291	0.0615	0.0886	-0.0105	0.2054
Novel 1	-0.1140	-0.0962	0.1340	-0.2719	-0.1924	0.0267
Novel 2	-0.2935	-0.2849	-0.0248	-0.3693	-0.2674	-0.0733
Novel 3	-0.3324	-0.3297	-0.0781	-0.3840	-0.2755	-0.1035
Coverage	0.2163	0.1896	0.0144	0.3369	0.2431	0.0688
Compression	-0.0879	-0.6086	-0.0041	-0.0697	-0.0084	-0.1155
Density	0.4591	0.4517	0.1991	0.4035	0.2738	0.2019

Table 3: Spearman correlation of our metrics and human judged quality metrics with various statistics describing summaries across the 1700 SummEval summaries.

ship between model size and average document info with and without the help of a summary. We can see that as the model gets larger, both average $I(\mathcal{D})$ and average $I(\mathcal{D}|\mathcal{S})$ decrease together. Larger models should be better at autoregressive token prediction, as reflected in the plot of $I(\mathcal{D})$, but it is interesting to see that $I(\mathcal{D}|\mathcal{S})$ decreases at around the same rate. We suspect this is because larger models may not be more suitable at utilizing a summary to predict a document under our setup.

5.2 Upstream Sentences

As described in section 2.3, we are making an independence assumption between sentences in a document when estimating $I(\mathcal{D})$, $I(\mathcal{D}|\mathcal{S})$, and $I(\mathcal{D}|\mathcal{D})$ by feeding each sentence into the model individually. We could alternatively assume that each sentence in the document is dependent on the k previous sentences, where $k = 0$ refers to our current approach and $k = \infty$ (or the maximum number of sentences in a document) drops the sentence independence assumption altogether. One could reason that this would better allow us to quantify the information in a document, which may lead to a more effective metric.

As shown in table 5, using $k > 0$ leads to an improvement in consistency at the expense of the other summary dimensions, and increasing k beyond 1 does not yield any significant gains in performance. Figure 4 shows that increasing k from 0 is more helpful at decreasing $I(\mathcal{D})$ than it is at decreasing $I(\mathcal{D}|\mathcal{S})$. We could draw a similar conclusion as we did in section 5.1 that increasing k is helpful for autoregressive token prediction, but it doesn't help our model with utilizing a summary to predict a document in our setup.

5.3 BLANC-Shannon

Our metrics bear similarity to the BLANC-help metric (Vasilyev, Dharnidharka, and Bohannon 2020; Vasilyev et al. 2020), which measures the accuracy of the BERT language model on the task of guessing masked tokens with and without a summary prepended to a document. The BLANC score is measured as a boost in unmasking accuracy $a_{help} - a_{base}$ when masking various sets of M evenly spaced tokens, where a_{help} is the accuracy when the summary is provided as help and a_{base} is the accuracy when no help is provided. Our metrics differ from BLANC in that we measure information instead of raw accuracy, we generate documents autoregressively instead of masking, and we typically use GPT-2 instead of BERT.

To study the utility of measuring document information as opposed to raw accuracy counts, we define BLANC-Shannon to be the boost in accuracy when generating document tokens given the summary. On the SummEval benchmark, BLANC-Shannon achieves Kendall tau-b system-level correlations of 0.3676, 0.6765, 0.5092, and 0.5588 for the expert annotations of coherence, consistency, fluency, and relevance respectively. These scores are an improvement on the consistency dimension over the Shannon Score and Information Difference metrics at the expense of every other dimension. We can only hypothesize that accuracy may be more sensitive to wrongly generated tokens and hence to consistency, but it would be interesting to compare BLANC-Shannon to the other metrics on an even larger dataset than SummEval.

6 Related Work

The Shannon Game The Shannon Game (Hovy and Lin 1998) was proposed over two decades ago as a way to use humans to measure the information retention between document and summary. In the original formulation, humans need to guess a document letter by letter given the summary, document, or nothing, and they measure the total number of guesses that were required to reconstruct the document. The authors ran a small-scale experiment where they conducted this game using human subjects, and they found a clear order of magnitude difference between the number of guesses each human required, as expected. However, they also found that reconstructing the original document with no help (the task of human 3) was extremely time-consuming, sometimes taking over 3 hours, making the Shannon Game prohibitively expensive as a human evaluation method.

Automated Summary Evaluation The most popular automatic summarization evaluation method is the ROUGE family of metrics (Lin 2004; Lin and Och 2004), which measure word overlap between the system summary and one or more reference summaries. The two biggest problems we see with ROUGE as a metric are 1) that it relies on human written reference summaries, and 2) that it measures simple word overlap, which means that a perfectly paraphrased version of the reference summary would score poorly.

Many solutions have been proposed to remedy issue #2 without solving issue #1, such as BERTScore (Zhang et al.

k	Coher.	Consi.	Fluen.	Relev.
		Shannon Score		
0	**0.4118**	**0.6324**	**0.5240**	**0.6029**
1	0.3529	**0.6618**	0.4945	0.5441
2	0.3235	**0.6618**	0.4945	0.5147
3	0.3235	**0.6618**	0.4945	0.5147
4	0.3235	**0.6618**	0.4945	0.5147
		Information Difference		
0	**0.4706**	**0.6324**	**0.5683**	**0.6618**
1	0.3529	**0.6618**	0.4945	0.5441
2	0.3382	**0.6765**	0.5092	0.5294
3	0.3235	**0.6618**	0.4945	0.5147
4	0.3382	**0.6765**	0.5092	0.5294

Table 5: Kendall tau-b system-level correlations between expert annotations of coherence, consistency, fluency, and relevance and our Shannon Score and Information Difference metrics with different choices of k (the number of upstream sentences to provide the model) on the SummEval dataset. Scores at least as high as those of $k = 0$ are bold. Coefficients above 0.36 are significant at the $\alpha = 0.05$ level.

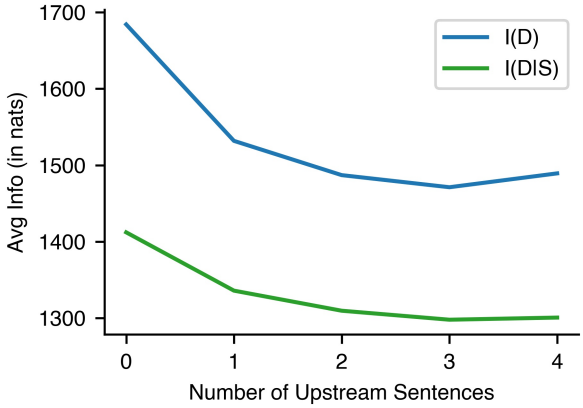

Figure 4: The average document information and document information given summary when prompting GPT-2 with different amounts of upstream sentences for the SummEval dataset.

2020), MoverScore (Zhao et al. 2019), Sentence Mover Similarity (Clark, Celikyilmaz, and Smith 2019), Word Mover Similarity (Kusner et al. 2015), and ROUGE-WE (Ng and Abrecht 2015). All of these metrics involve the idea of using soft overlap or embedding/token distance between the system and reference summaries. Louis and Nenkova (2009) suggested measuring the Jensen-Shannon divergence between word distributions used in the system summary and original document, which suffers from issue #2 while fixing issue #1. Sun and Nenkova (2019) and Gao, Zhao, and Eger (2020) perform reference-free summary evaluation using language model word embeddings with promising results. Other have used question generation and question answering models to evaluate summaries (Scialom et al. 2019; Chen et al. 2018), but we argue that these metrics are only as good as the datasets the models were trained on, and may have problems generalizing. Beyond summarization, there have been many metrics proposed for Natural Language Generation more generally (Sai, Mohankumar, and Khapra 2020).

Our methods are most similar to BLANC (Vasilyev, Dharnidharka, and Bohannon 2020; Vasilyev et al. 2020), which measures the accuracy boost of BERT (Devlin et al. 2019) on the Cloze task (Taylor 1953) when a summary is prepended to a document or the model is finetuned on the summary. This paper contributes to the study of BLANC-like metrics by extending them to new language models, giving them a theoretical motivation, and performing more robust experiments to better understand their behavior. The information-theoretic motivations of our metrics are similar to that of Peyrard (2019) who formally defined some metrics based on distributions of semantic units, which contrasts with our use of pretrained language models.

7 Conclusion

In this work, we successfully show that a universal language model performing the basic language modeling task is an effective reference-free evaluator of summary quality. This work extends the Shannon Game from using humans as evaluators to using machines, and extends the work on BLANC-like metrics to new language models and theoretical interpretations. We experimentally showed that our metrics strongly correlate with expert judgement of summary quality, and hope that they will serve as useful tools for the future development of summarization models. As next steps, it would be interesting to see if our metrics are useful for summarization model training, or evaluation in tasks beyond standard summarization, such as paraphrasing or query-focused summarization. Our code is available on GitHub.[1]

References

Chen, P.; Wu, F.; Wang, T.; and Ding, W. 2018. A Semantic QA-Based Approach for Text Summarization Evaluation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 4800–4807. AAAI Press (2018).

Clark, E.; Celikyilmaz, A.; and Smith, N. A. 2019. Sentence Mover's Similarity: Automatic Evaluation for Multi-Sentence Texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2748–2760. Florence, Italy: Association for Computational Linguistics.

Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.; and Salakhutdinov, R. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2978–2988. Florence, Italy: Association for Computational Linguistics.

[1] github.com/primerai/blanc/tree/master/shannon

Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; and Liu, R. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *International Conference on Learning Representations*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Fabbri, A. R.; Kryściński, W.; McCann, B.; Xiong, C.; Socher, R.; and Radev, D. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9: 391–409.

Gao, Y.; Zhao, W.; and Eger, S. 2020. SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1347–1354. Online: Association for Computational Linguistics.

Grusky, M.; Naaman, M.; and Artzi, Y. 2018. Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 708–719. New Orleans, Louisiana: Association for Computational Linguistics.

Hermann, K. M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching Machines to Read and Comprehend. In Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 28, 1693–1701. Curran Associates, Inc.

Hovy, E.; and Lin, C.-Y. 1998. 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*, 197–214. Baltimore, Maryland, USA: Association for Computational Linguistics.

Keskar, N. S.; McCann, B.; Varshney, L.; Xiong, C.; and Socher, R. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From Word Embeddings To Document Distances. volume 37 of *Proceedings of Machine Learning Research*, 957–966. Lille, France: Proceedings of Machine Learning Research (PMLR).

Lavie, A.; and Agarwal, A. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, 228–231. Prague, Czech Republic: Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Lin, C.-Y.; and Hovy, E. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 150–157.

Lin, C.-Y.; and Och, F. J. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 605–612. Barcelona, Spain.

Louis, A.; and Nenkova, A. 2009. Automatically Evaluating Content Selection in Summarization without Human Models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 306–314. Singapore: Association for Computational Linguistics.

Louis, A.; and Nenkova, A. 2013. Automatically Assessing Machine Summary Content Without a Gold Standard. *Computational Linguistics*, 39(2): 267–300.

Ng, J.-P.; and Abrecht, V. 2015. Better Summarization Evaluation with Word Embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1925–1930. Lisbon, Portugal: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 311–318. USA: Association for Computational Linguistics.

Peyrard, M. 2019. A Simple Theoretical Model of Importance for Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1059–1073. Florence, Italy: Association for Computational Linguistics.

Peyrard, M.; Botschen, T.; and Gurevych, I. 2017. Learning to Score System Summaries for Better Content Selection Evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, 74–84. Copenhagen, Denmark: Association for Computational Linguistics.

Popović, M. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392–395. Lisbon, Portugal: Association for Computational Linguistics.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving Language Understanding by Generative Pre-Training. OpenAI.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. OpenAI.

Sai, A. B.; Mohankumar, A. K.; and Khapra, M. M. 2020. A Survey of Evaluation Metrics Used for NLG Systems. arXiv:2008.12009.

Scialom, T.; Lamprier, S.; Piwowarski, B.; and Staiano, J. 2019. Answers Unite! Unsupervised Metrics for Reinforced

Summarization Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3246–3256. Hong Kong, China: Association for Computational Linguistics.

Sun, S.; and Nenkova, A. 2019. The Feasibility of Embedding Based Automatic Evaluation for Single Document Summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1216–1221. Hong Kong, China: Association for Computational Linguistics.

Taylor, W. L. 1953. Cloze procedure: A new tool for measuring readability. *Journalism Bulletin*, 30(4): 415–433.

Vasilyev, O.; Dharnidharka, V.; and Bohannon, J. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, 11–20. Association for Computational Linguistics.

Vasilyev, O.; Dharnidharka, V.; Egan, N.; Chambliss, C.; and Bohannon, J. 2020. Sensitivity of BLANC to human-scored qualities of text summaries. *arXiv preprint*, arXiv:2010.06716.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zhao, W.; Peyrard, M.; Liu, F.; Gao, Y.; Meyer, C. M.; and Eger, S. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 563–578. Hong Kong, China: Association for Computational Linguistics.

Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. *2015 IEEE International Conference on Computer Vision (ICCV)*, 19–27.