ShapSum: A Framework to Predict Human Judgement Multi-Dimensional Quality Scores for Text Summarization

Carolina Arriaga U.C. Berkeley
Ayman Moawad U.C. Berkeley
Abhi Sharma U.C. Berkeley
[caro.arriaga,aymoawad,abhisha]@berkeley.edu

Abstract

Text summarization is the task of producing a shorter version of a document. Model performance has been compared amongst each other based mainly on their ROUGE score. The metric has been widely criticized because it only assesses content selection and does not account for other quality metrics such as fluency, grammaticality, coherence, consistency, and relevance (Ruder), (Lin, 2004). Combined score metrics like BLEND or DPMFcomb incorporate lexical, syntactic and semantic based metrics and achieve high correlation with human judgement (Yu et al., 2015) in the MT and text generation context. However, none of these combined metrics have been tested in summaries nor have moved away from human scores based on Pyramid and Responsiveness scores. Our findings show that multiple metrics used in the summarization field are predictive of multidimensional quality evaluations from experts. We produced four saturated models using decision trees and the corresponding surrogate Shapley explanation models to measure metric contribution against four dimensions of evaluation (fluency, relevance, consistency, coherence). We hope that our work can be used as a standard evaluation framework to compare summary quality between new summarization models.

All related code and data can be found here: ShapSum

1 Introduction

Text summarization is the task of producing a shorter version of a document. Humans have the natural ability to understand what components of a text are the most important and remove secondary information. The current state of the art in summarization tasks is divided into two types of models: extractive and abstractive. Extractive models focus mainly on sentence ranking like LEAD (Narayan et al., 2018) or selection-compression models that take two steps in the summarization process (Xu and Durrett, 2019). Extractive systems are robust and straightforward to use (Grusky et al., 2018). Abstractive systems seek to conceptualize a text by utilizing context architectures that use encoder-decoder architectures like BERT (Devlin et al., 2019) or autoregressive models that use an attention mechanism such as GPT (Radford et al., 2018).

Model performance has been compared amongst each other based mainly on their ROUGE score (Lin, 2004), a recall-oriented metric that counts the number of overlapping units (n-gram) between golden and decoded summaries. Although ROUGE has been a widely used metric due to its high correlation with human judgments, it has been also widely criticized because it only assesses content selection and does not account for other quality metrics such as fluency, grammaticality, coherence, consistency, and relevance (Ruder).

Multiple metrics appeared to evaluate the aforementioned dimensions, some requiring golden references while others are referenceless. These metrics are compared with human judgment scores from datasets created as by-products of the manual evaluations performed during the DUC[1]/TAC[2] shared tasks. Current summarization systems now operate within the higher-scoring range where fewer data is collected and judgment scores strongly disagree (Peyrard, 2019). This is problematic because there's is a lack of high score summaries to evaluate these novel architectures. SummEval its a toolkit which incorporates both, metrics and human annotations, in a consistent and comprehensive fashion (Fabbri et al., 2021). We selected a subset of the 14 metrics they propose that correlate well with human judgments and extended the metric space. We build upon this line of research, seeking to understand the relationship between four quality dimension for human judge-

[1] http://tac.nist.gov
[2] Text Analysis Conference

ments (coherence, consistency, fluency, and relevance) and highest correlated metrics. Our main goals are:

- To find out whether multiple metrics widely used in the summarization field are predictive of multidimensional quality evaluations from experts.
- To generate a model that generalizes well with human annotation scores within the CNN/DM dataset regardless of the system used to produce the summary.
- To provide a framework that facilitates the scoring in each four dimensions of quality.

2 Related work

The progress in summarization is highly tied to a few metrics that remain as the baseline for quality. These metrics, like ROUGE and its variations, have shown to correlate well with human judgements (Lin and Hovy, 2003), especially when extractive models are used. Abstractive models introduce novel n-grams to a summary, therefore overlapping n-grams are unable to rate a model's quality. Modern systems provide high quality summaries (Pegasus, T5) while most of the annotated data collected in DUC/TAC conferences tend to generate average summary scores (Peyrard, 2019) and usually take time to release. The last DUC dataset was produced in 2007. Researchers are forced to introduce new metrics and compare it with old annotation datasets. An estimate of twenty three different metrics have been use to evaluate summaries in the last decade. For example, BERTScore metric does not search for exact matches, it computes the token similarity using contextual embeddings (Zhang et al., 2019). This scoring method brings flexibility to the a model's evaluation. A similar concept is applied to ROUGE-WE (word embedding) metrics that extends ROUGE by using soft lexical matching based on the cosine similarity of Word2Vec embeddings. SummaQA applies a BERT-based question-answering model to answer cloze-style questions using generated summaries (Scialom et al., 2019). The comparison of new metrics with human judgement ranges from application and experimental setup. BERTScore showed a high correlation with human judgements for a Machine Translation (MT) task between multiple languages. MoverScore, a metric that creates combination of contextualized representations of system and texts in conjunction with their semantic distance (Zhao et al., 2019), had a high correlation with Pyramid score (Nenkova et al., 2007) and Responsiveness score from TAC. Under these different experimentation settings, and generalizing to multiple metrics, we derive that each of them cover an isolated quality dimension within human judgements. Combined score metrics like BLEND or DPMFcomb incorporate lexical, syntactic and semantic based metrics and achieve high correlation with human judgement (Yu et al., 2015) in the MT and text generation context. However, none of these combined metrics have been used to test summaries nor have moved away from the correlation with the Workshop on WMT[3] annotated scores.

3 Data

3.1 CNN Daily-Mail

A standard dataset for summarization tasks is the CNN/DailyMail corpus (Hermann et al., 2015), originally a question answering task, which was repurposed for summarization by (Nallapati et al., 2016). The dataset consists of news articles and associated human-created bullet-point summaries. In all, the corpus has 286,817 training pairs, 13,368 validation pairs and 11,487 test pairs, as defined by their scripts. The source documents in the training set have 766 words spanning 29.74 sentences on an average while the summaries consist of 53 words and 3.72 sentences.

3.2 Human Judgement Annotations

The authors of SummEval (Fabbri et al., 2021) annotate summaries produced by 16 different models for the same 100 articles, resulting in a total of 1600 annotated summaries. The 100 chosen articles are randomly picked from the test corpus of CNN-DailyMail and are each run through 16 summary models (consisting of both abstractive and extractive techniques). Each of the 1600 summaries are then annotated on different human summary quality dimensions (coherence, fluency, consistency, relevance) on a likert scale from 1-5. These summaries are annotated by 3 independent expert judges and 5 independent high reputation annotators from Mechanical Turk. We evaluate our summarization metrics against these 1600 summary outputs and we attempt to model the relationship between these metrics and the mean quality score of interest. Only

[3]Statistical Machine Translation

expert scores are selected for consideration because expert scores are shown to provide different judgements against turker judgements for the same summary and have less inter-annotator disagreement (Gillick and Liu, 2010). A summary of the expert scores across all quality dimensions has been shown in table 1

Table 1: All expert judge annotation statistics (Median and Mean)

	Mean	Median
Coherence	3.42	3
Relevance	3.78	4
Consistency	4.66	5
Fluency	4.67	5

4 Methods

We introduce the universe of metrics and the modeling approach of a combined score.

4.1 Evaluation Metrics

The selection of evaluation metrics includes primarily metrics based on the SummEval benchmark. Correlations reported by Fabbri et al. (2021) don't include three unsupervised metrics used for summarization which will be briefly described below.

SLOR/WPSLOR (Kann et al., 2018), syntactic log-odds ratio is a normalized language model score, as a metric for referenceless fluency evaluation of natural language generation output at the sentence level. The former expands into word piece tokenization and is the metric used in our model.

Topic Similarity, based on Latent Dirchlet Allocation model (Blei et al., 2003), is the cosine similarity between two topic distribution vectors.

Entity Grid(Barzilay and Lapata, 2008), is the entity-grid representation of discourse, which captures patterns of entity distribution in a text. The algorithm introduced in the article automatically abstracts a text into a set of entity transition sequences and records distributional, syntactic, and referential information about discourse entities.

4.2 Predictive Modeling and Shapley

We want to evaluate summaries against the four dimensions of human-level expert evaluation (fluency, relevance, consistency, coherence) with respect to their metric scores. Different models generate different summaries of different qualities according to human evaluation scores. When human evaluation is not possible/available we would like a combination of metric scores to allow to predict human annotation. In the following we aim to model expert evaluation scores given calculated metrics in a typical discriminative supervised learning setting. Given a dataset $\mathcal{D} = \{(\boldsymbol{X_i}, y_i)\}_{i=1,...n}$ of n summaries, we have $\boldsymbol{X_i} \in \mathbb{R}^p$ metric scores and $y \in \mathbb{R}^m$ expert evaluations, we want to learn a function $f : \boldsymbol{X} \mapsto y$ that can predict y^* for a new, unobserved document with calculated metric score inputs $\boldsymbol{X^*}$.

We leverage state-of-the art gradient boosting on decision tree algorithms as a popular machine learning model for data in tabular form with heterogeneous features (Chen and Guestrin, 2016). It is designed to achieve competitive results in the presence of complex, noisy and highly feature-dependent data. Xgboost like many other complex machine learning models, is very flexible and is able to model non linearities and complex feature interaction. However, this makes the statistical inference challenging. In addition to predictive power, given the model output $f(x_1, \ldots, x_m)$, one would want to quantify to what extent each x_j contributes to the expert evaluation score output. In order to relate the input metrics to the predictions without loss of model accuracy (i.e. avoiding a pure linear model approach), we can relate the input features to the predictions in a non linear complex setting on the basis of a coalitional game theory method using the computation of Shapley values (Shapley, 1953). Shapley values allow to fairly allocate credits to features of a model for the output of that model. It has strong theoretical support to ensure a fair feature attribution and consequently a fair distribution of the total expert annotation prediction value among the metrics and their individual contributions.

The explanation of complex models via Shapley values starts by defining a class of additive feature attribution methods that will be used as a surrogate explanation model for the original one. If f is the original prediction model and g is an explanation model, then an additive feature attribution model is a linear function of binary variables in the form:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i$$

where M is the number of input features, $z' \in \{0,1\}^M$ are the features being observed or unob-

served, respectively, and $z'_i = 1$ or $z'_i = 0$ and $\phi_i \in \mathbb{R}$ are the feature attribution values.

Given a model prediction $f(x)$, by assigning a feature mapping function $h_x(z')$ that maps binary inputs to the original feature space such that $x = h_x(z')$, we can evaluate $f(h_x(z'))$ and calculate the effect of observing or not observing a feature and seek to enforce $f(x) = f(h_x(z')) \approx g(z')$ through a special selection of ϕ_i. Shapley, through his work, described desirable properties constraining the space of solutions for ϕ_i:

- **Local accuracy/additivity/efficiency.** The sum of feature attributes need to match the original model output.
- **Missingness.** If a feature is missing, it receives zero attribution.
- **Consistency/monotonicity.** For two different models f_1 and f_2 in the same feature space, if the contribution of a feature i increases for f_2 vs. f_1, then the given attribution for feature i should not decrease for f_2.
- **Symmetry.** If i and j are two features that contribute equally, their attribution should be equal.
- **Linearity.** The attributions of the sum of two functions f_1 and f_2 expands to the sum of the attributions for each of the two functions.

Those mathematically axiomatized properties (see (Lundberg and Lee, 2017) for details) describe a *fairness* context of attribution.

Let $S \subseteq \mathcal{M} = \{1, \ldots, M\}$, a subset of non-zero indexes. By defining $f_x(S) = f(h_x(z')) = \mathbb{E}[f(x)|\mathbf{do}(x_S)]$ then the only set of values (Shapley, 1953), (Young, 1985) for the explanation model satisfying the above properties can be proven to be:

$$\phi_i(f, x) = \sum_{S \subseteq \mathcal{M} \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} \Big[f_x(S \cup \{i\}) - f_x(S) \Big] \tag{1}$$

The above quantity represents some form of weighted average of the assigned attributions, calculated from model evaluation difference with and without the feature of interest, over all possible subsets of features S. By computing shapley values we can extract partial dependencies and as a results infer on the effect of each metric on the expert annotation scores.

4.3 Quality Score Prediction - Analysis Across Varying Length Summaries

In addition to our main analysis, we perform an analysis on the relationship between summary length and quality scores. The authors of the SummEval paper only obtained unconstrained length summaries from their models and annotated them. We believe that constraining summary length can have an impact on quality scores. Since this is an auxiliary experiment to our main analysis, we don't report the results for it in this paper. However, if the reader is interested in learning more, we encourage them to learn more in our repository: ShapSum. We attempt to use the most important metrics from this paper that predict a quality dimension (say Fluency), and we use that to plot fluency scores for varying summary lengths for summaries produced by different models.

5 Results and Discussion

In this section we show an analysis of document summary metrics scores and their relationship to expert evaluation leveraging Shapley attribution method. In table 2 we present key model performance results on a test set of 20% of the summaries held for evaluation. We note that fluency has the best prediction performance with an average prediction error of roughly 0.6 points. On the flip side, the coherence scores are harder to predict with an average error of approximately 0.8 points.

Table 2: Test set model performance statistics

	RMSE	MAE
Coherence	0.86	0.82
Relevance	0.69	0.71
Consistency	0.79	0.64
Fluency	0.59	0.60

5.1 Relevance

Relevance refers to the selection of important content from the source i.e a summary that includes only important information from the original source article. In this sub-section, we attempt to analyze the contribution of each metric on the human annotation relevance score by extracting the corresponding Shapley values.

At the individual article summary level, figure 1 presents an example of walk-through of the individual metric scores and their contribution to the relevance score for the summary of Pegasus

model on article 1 (see appendix A). This summary scores high on relevance according to human annotation with a score of 4.3 averaging over all experts. Our model is able to predict a relevance score of 4.2. The path to prediction shows the metric feature values and their effect on the relevance score with respect to $E(f(X))$ the average relevance score across all articles (Shapley values are computed in relation to a reference baseline relevance score represented by the average article). We note that `rouge_we_1_f1` and `bert_score_recall` metric scores and their corresponding feature values had the most effect on attributing a high relevance annotation.

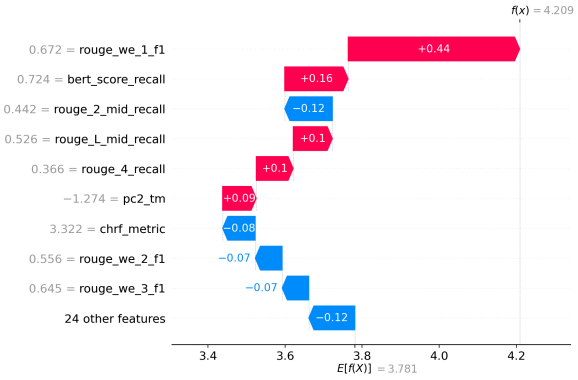


Figure 1: Example of relevance prediction with the contribution values of individual metrics towards relevance score for article 1.

Through the computation of Shapley attribution values for all the articles, and because every article will have an distinct attribution value for each of its metrics, we can aggregate all the articles and focus on one metric at a time. We can understand on a global level, the overall effects that metrics have on relevance. In figure 2, we show how individual metrics dependency plots can extract metric-relevance score relationships by looking at the attributed Shapley value against the value of the metric of interest. This relationship shows how a feature attribution changes as the feature value varies. The figure shows `rouge_we_1_f1` dependency on the relevance score. We observe a mostly linear (although somewhat S-shaped) positive relationship between the `rouge_we_1_f1` score and the resulting human annotation relevance score. We also note that a `rouge_we_1_f1` score of 0.45 and beyond is a clear cut-off for positive Shapley values as positive Shapley values have a positive effect on moving the relevance score upward with respect to the average relevance score. We also note that an increase in `rouge_we_1_f1` metric has a linear impact on relevance up to a value of roughly 0.55 after which the effect is constant.

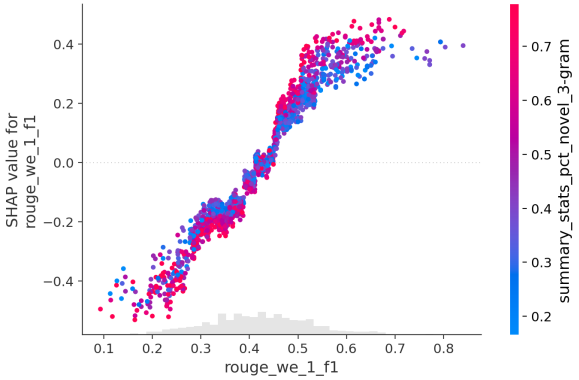
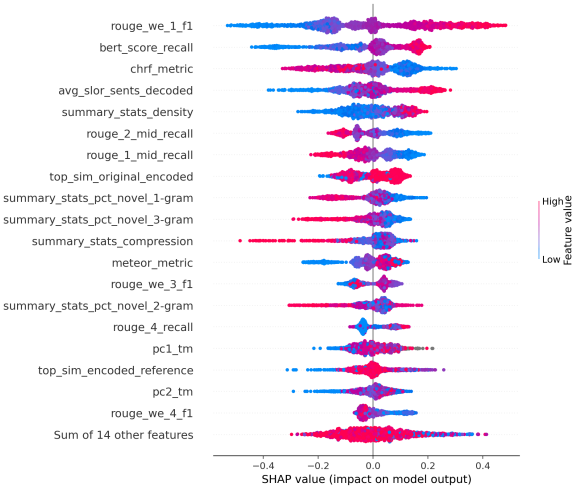


Figure 2: Dependency plot of the relationship between `rouge_we_1_f1` attributed relevance and the feature value. Every dot is a summary's attributed Shapley value towards relevance against the metric score feature value. The color mapping shows the most interacting metric with `rouge_we_1_f1` and its effect on dispersing the Shapley score as its value changes and interacts.

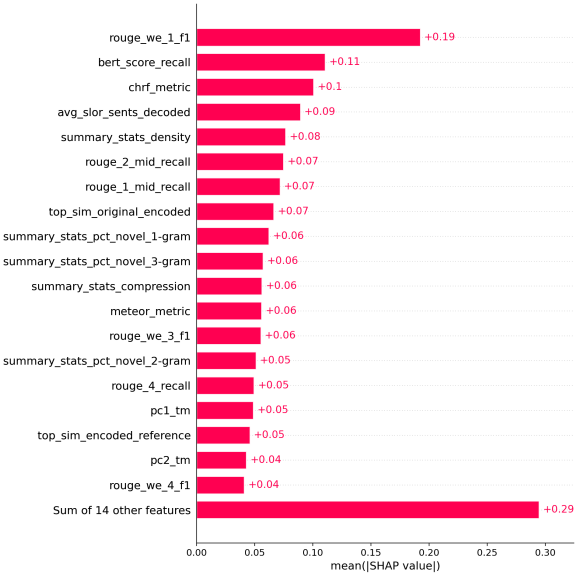
The Shapley values can be used to identify the importance of each features to the human annotated relevance score. Metrics have the most impact when the change in the model output (i.e the relevance) is greatly affected by the metric value. For linear models $f(x) = x^T\beta$, the coefficients of the covariates provide some clues. Alternatively, because of Shapley's natural local property, the attribution provided by Shapley derivation gives an individualized feature importance measure for each prediction over each metric. Their aggregation can ultimately provide an equivalent global importance measure, but the natural decomposition produces a richer view of importance and delivers higher resolution plots. In addition, Shapley solutions ensure consistency as stated in section 4.2.

Figure 3a shows the individual Shapley attribution values for a subset of the top metrics that contribute the most to relevance. High Shapley values mean a high relevance attribution, which depends on the feature value shown by the color code. The plot gives a high resolution feel for feature importance, as each dot is a summary "feature attribution" value. The amplitude provides a general idea of the overall distribution over Shapley values that each metric has. The metrics are ordered by order of importance by summing over the N summaries j, i.e., $\frac{1}{N}\sum_{j=1}^{N}|\phi_i(f, x_j)|$ for each metric $i \in \mathcal{M}$. Figure 3b shows a standard feature importance bar

chart which measures global importance through summation over all summaries. For example, the figure shows that the `rouge_we_1_f1` is the most influential metric affecting relevance scoring. The higher the Shapley value, the bigger the contribution to relevance, and from the colormap in 3a we see that as `rouge_we_1_f1` metric value increases human annotation relevance increases. The large variance also provides information on the spread of the metric values in the dataset, and the density of dots shows how common each value of the metric is across all summaries.



(a)



(b)

Figure 3: (a) Individual (one dot per article) Shapley attribution values for top 20 metrics. High Shapley values mean a relevance attribution, which depends on the metric value shown by the colomap. (b) Standard feature importance bar chart for relevance score.

5.2 Coherence

Coherence refers to the collective quality of all sentences. The summary should build from sentence to sentence to a coherent body of information about a topic (Dang, 2005). Correlations reported by Fabbri et al. (2021) showed weak correlations on the coherence dimension. For this reason, we incorporated WPSLOR, topic similarity and two principal components taken from a transition matrix using the EGrid method (Barzilay and Lapata, 2008). After running the experiments, we found that the most important contributors to this dimension were `summary_stats_density`, `rouge_we_1_f1`, `bert_score_recall`, `slor_avg_sentences`.

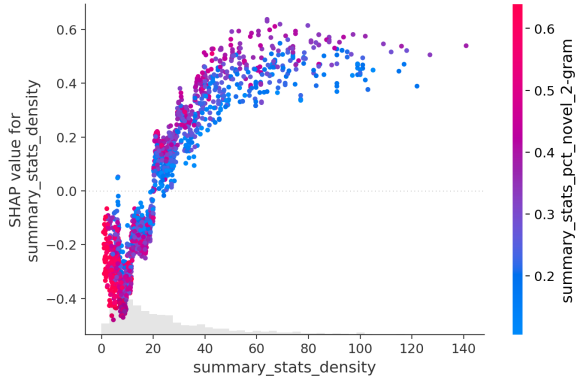


Figure 4: Dependency plot of the relationship between summary stats density and its SHAP values. Higher density leads to higher contribution to coherence and SHAP value levels off at around 0.4

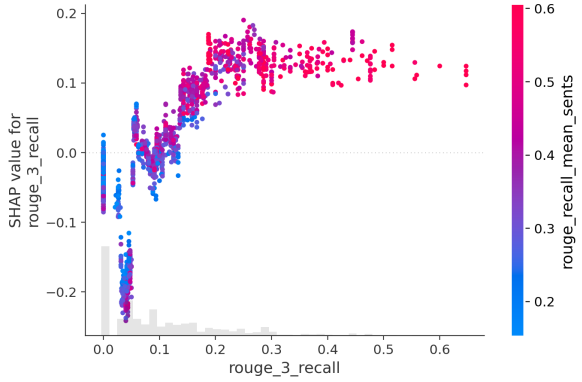


Figure 5: Higher order ROUGE like ROUGE-3 shows to contribute positively to coherence. This makes sense because higher overlap with golden summary should lead to coherent text.

Sentences that were longer than 20 words tend to create more coherent texts as seen in figure 4. The coherence contribution becomes stable af-

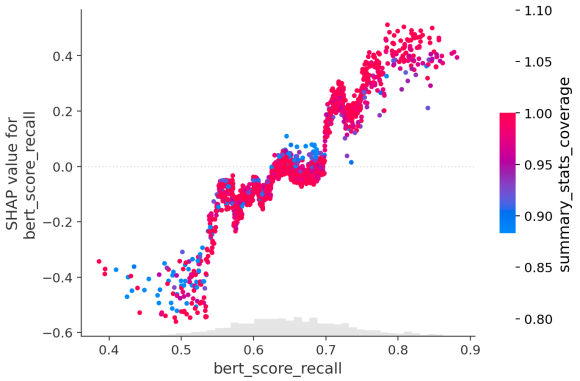


Figure 6: Contextualized embedding score metrics like BERT Score are positively impacting coherence beyond 0.65 threshold.

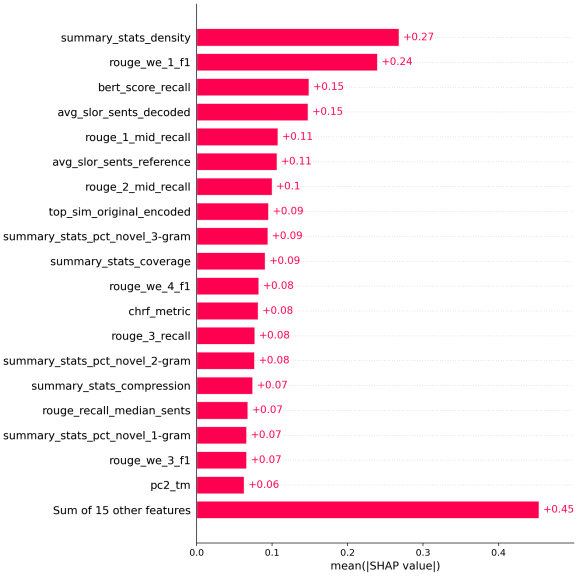


Figure 7: Standard feature importance bar chart for coherence score.

ter reaching 40 words, where it reaches a maximum value of 0.4 points. The `rouge_we_1_f1` behaves identical to figure 2, except that the average contribution to the coherence score is smaller than `summary_stats_density`. Overall, metrics contribution can be seen on figure 7. Both `bert_score_recall` and `slor_avg_sents_decoded` have an equal impact on coherence, and have a positive high correlation reaching a maximum SHAP score of 0.4 value.

These findings suggest that the most important metrics for coherence rely on characteristics like: sentence length, semantic similarity of words (better for abstractive models), contextual embeddings and likeliness of a sentence based on a pre-selected language model. For our purposes, we selected the language model dictated by GPT-2 medium model to report SLOR scores. Interestingly, the EGRID metric did not reveal too much overall contribution in coherence. Both, the topic similarity metric as well as SLOR did rank in the top 10 contributors to coherence based on figure 7. So we are encouraged by the fact that our novel metrics did provide predictive signal for coherence.

5.3 Fluency

Fleuncy is scoring sentence level quality. This score is penalizing sentences in the summary that have formatting problems, are incomplete, are ungrammatical or have missing text. The highest scoring metrics in this domain are shown in Figure 8, but overall - the metrics to highlight are BERTScore recall, `summary_stats_density` and `avg_slor_sents_decoded`. Upon investigating the effect of `bert_score_recall`, we observe that its effect on SHAP seems to be roughly constant, around 0.1 after obtaining a BERTscore recall of 0.65 (or above) on a given summary. Notice how - similar to coherence, a BERTScore recall of 0.65 and above leads to both coherent and fluent summaries.

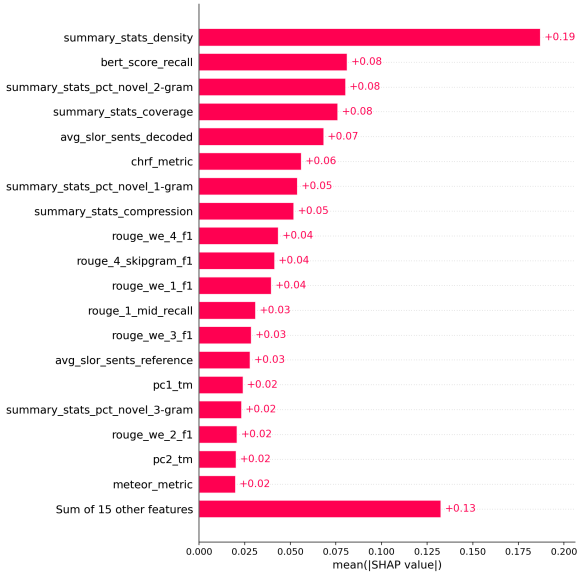


Figure 8: Standard feature importance bar chart for fluency score.

The more interesting relationships are highlighted in metrics below. Though some of these effects are not very large, they do show a trend, which is worth investigating. For example, `avg_slor_sents_decoded` in Figure 9 shows that as SLOR value increases (less negative is more

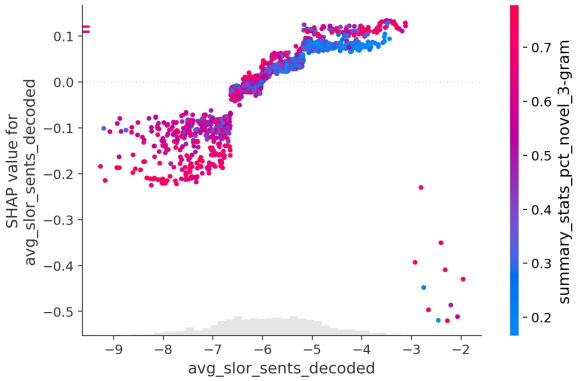


Figure 9: SLOR values highlighting effect on sentence level fluency. Less negative implies higher sentence fluency. Scores are reported as the average of all SLOR sentence scores in the summary.

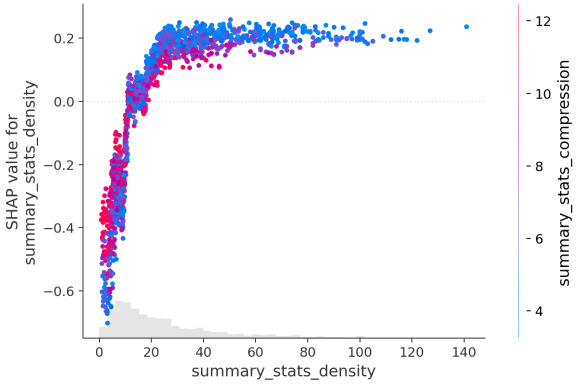


Figure 10: As sentence density increases, we observe higher impact on fluency. The effect on SHAP levels off after density reaches 30, which implies that a sentence with density=30 is approx. as fluent as sentence with density=40.

fluent), we observe a positive effect on SHAP values contributing to fluency. In Figure 10, we observe that the effect of density on SHAP levels off after reaching a threshold length of 30.

5.4 Consistency

This score checks for the quality of factual alignment between the summary and the article. It will penalize summaries that "dream up" facts that don't exist in the article. We didn't find any interesting strong correlations in the graphs for this dimension. The only thing we did find was that generating novel n-grams gets penalized from the point of view of generating consistent summaries. This makes sense because creating novel n-grams, i.e., new text is correlated with "making up" text that may not entail the source article. Thus, we see that for small values of novel-ngram, the model very slightly encourages creation of new n-grams (positive impact on SHAP scores). After a certain threshold (0.5 for bi-grams and 0.6 for tri-grams), the effect on SHAP decreases sharply, indicating negative effect on consistency. This can be seen in Figure 12 and 13

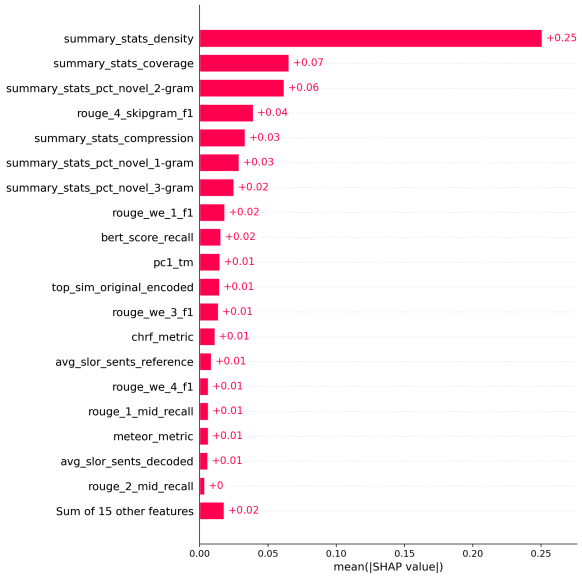


Figure 11: Standard feature importance bar chart for consistency score.

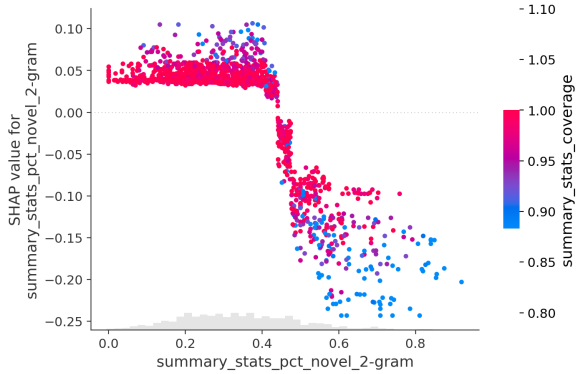


Figure 12: After crossing a threshold, the penalty on SHAP for consistency scoring increases sharply. Threshold for bi-gram is 0.5

6 Comparing Summaries

The Shapley method allows to compare summaries and their four dimensions of evaluation (fluency, relevance, consistency, coherence) with respect to their metric scores, allowing to provide an explanatory framework for every evaluation. Different models generate different summaries of different qualities according to human evaluation scores. When human evaluation is not possible/available, a

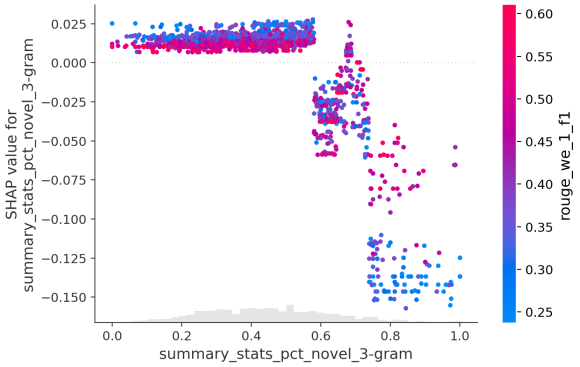


Figure 13: After crossing a threshold, the penalty on SHAP for consistency scoring increases sharply. Threshold for tri-gram is 0.6

combination of metric scores allows to predict human annotation as shown in the previous sections. It is of interest to provide insight in regards to what metric contributed to positively/negatively to the final evaluation on 2 (or more) summaries. In Figure 14, we compare the *predicted* consistency scores of two summaries for two different models. Below are the two summaries, with their true consistency scores, and "in red" we highlight sources of hallucination and inconsistencies with the original text:

- **Model 1 - True Consistency score=1.0:** paul merson was brought on with only seven minutes remaining in his team 's 0-0 draw with burnley . andros townsend scored the tottenham midfielder in the 89th minute . paul merson had another dig at andros townsend after his appearance . the midfielder had been brought on to the england squad last week . click here for all the latest arsenal news news .

- **Model 2 - True Consistency score=5.0:** paul merson has restarted his row with andros townsend . the tottenham midfielder was brought on with only seven minutes remaining in his team 's 0-0 draw with burnley . andros townsend scores england 's equaliser in their 1-1 friendly draw with italy in turin .

In Figure 14 we note that the predicted consistency scores are $\approx$ 1.7 and $\approx$ 4.8. The difference in metric values that explain divergence between the two scores are attributed to the top few metrics and their feature values (not shown here). We note that `summary_stats_density` has the largest slope among all metrics, showing a great impact on consistency scoring. For reference, `summary_stats_density=8.015` for Model 1 and `summary_stats_density=16.2` for Model 2.

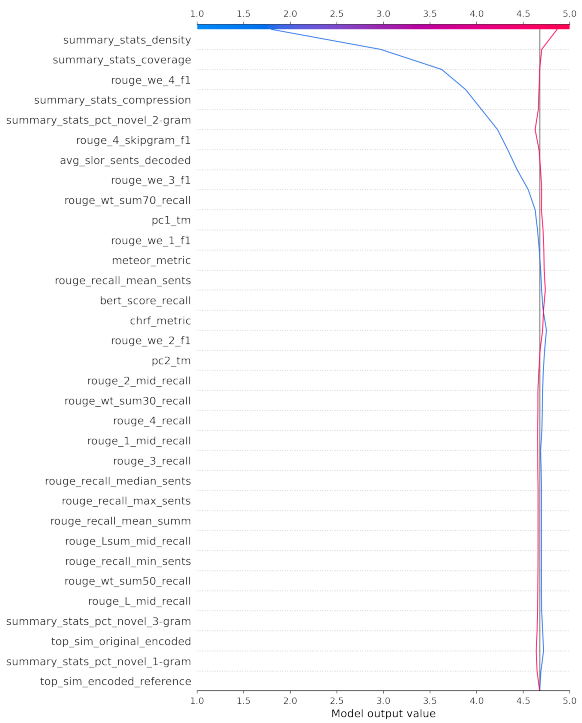


Figure 14: Example of two summaries consistency predictions and how metric scores feature value affect the prediction result.

7 Conclusions

In this paper, we built upon the research done by the SummEval team. We quantified the effect of summarization metrics on human judgement quality dimensions (consistency, fluency, relevance, coherence) and we built a framework to analyze, predict and compare summaries with these metrics, isolating them by dimensions of interest. We also introduced 3 novel metrics for coherence and we showed how 2 of these metrics contributed positive signal to coherence. Lastly, we performed an auxiliary analysis to show the impact of varying summary lengths on these quality dimensions.

Acknowledgments

We would like to thank the instructors of MIDS W266 at U.C. Berkeley - Prof. Joachim Rahmfeld, Prof. Daniel Cer and Prof. Mark Butler and others for providing us regular guidance throughout our program and being generous with their time and support. We would also like to thank the TAs for

helping grade assignments and answer questions for us regularly.

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:2003.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pages 785–794. ArXiv: 1603.02754.

Hoa Trang Dang. 2005. Overview of duc 2005. In *In Proceedings of the Document Understanding Conf. Wksp. 2005 (DUC 2005) at the Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, Los Angeles. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! *arXiv preprint arXiv:1809.08731*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chin-Yew Lin and Eduard Hovy. 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*, pages 150–157.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2):4–es.

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

S. Ruder. Summarization.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. *arXiv preprint arXiv:1909.01610*.

Lloyd S. Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317.

Jiacheng Xu and Greg Durrett. 2019. Neural extractive text summarization with syntactic compression. *arXiv preprint arXiv:1902.00863*.

H. P. Young. 1985. Monotonic solutions of cooperative games. *International Journal of Game Theory*, 14(2):65–72.

Hui Yu, Qingsong Ma, Xiaofeng Wu, and Qun Liu. 2015. Casict-dcu participation in wmt2015 metrics task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 417–421.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

A Article 1

original: Priscilla Presley will serve as a witness at the first wedding to be held at an all-new chapel of love in Las Vegas. The 69-year-old collaborated with NBC's Today show to launch a contest for one Elvis-obsessed couple to win the 'ultimate wedding'. The winning duo - announced next Monday - will tie the knot at Elvis Presley's Graceland Wedding Chapel inside the Westgate Hotel on Thursday, April 23. Novel idea: Priscilla Presley will serve as a witness at the first wedding to be held at an all new chapel of love in Las Vegas . Westgate, formerly the Las Vegas Hilton, is where Elvis performed more than 830 sold-out shows. Along with the singer's former wife in the audience, the winning couple will win a free wedding reception and hotel suite for two nights. To top it off, airfares and concert tickets to the Elvis Experience theater show will also be thrown in. While Priscilla agreed to make an appearance, the woman who wed Elvis in 1967 made one thing clear before unveiling the latest wedding chapel to bear his name: No impersonators. 'This is all first-class,' she told the Associated Press recently. 'This is not a joke. The wedding chapel is not a joke.' The actress and business magnate has been involved in the look and feel of the chapel that officials say is part of the first permanent Graceland exhibit of the singer's artifacts outside his iconic Memphis, Tennessee, home. Couples wanting to be the first to wed at the Elvis Presley's Graceland Wedding Chapel must submit a video or photos along with an explanation detailing why they deserve the prize. Chapel of love: The winning duo - announced next Monday - will get married in the brand new Elvis Presley's Graceland Wedding Chapel (artist's rendering above) at the Westgate Hotel on Thursday, April 23 . Flashback: In this May 1, 1967, file photo, singer Elvis Presley, 32, and his bride, 21, the former Priscilla Beaulieu, appear at the Aladdin Hotel in Las Vegas, after their wedding .

Pegasus summary: the 69-year-old actress collaborated with nbc's today show to launch a contest for one elvis-obsessed couple to win the 'ultimate wedding' the winning duo - announced next monday - will tie the knot at elvis presley's graceland wedding chapel inside the westgate hotel on thursday, april 23 .