ShapSum: A framework to predict human judgement multi-dimensional qualities for text summarization.

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

Abstract

Insert abstract

1 Introduction

Text summarization is the task of producing a shorter version of a document. Humans have the natural (learned) ability to understand what components of a text are the most important and remove information that is secondary to the main idea. 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., 2018) 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 human 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 have appeared to evaluate the aforementioned dimensions, some requiring golden references while others are referenceless [citation needed]. These metrics are typically compared with human judgment datasets created as by-products of the manual evaluations performed during the DUC/TAC shared tasks which now strongly disagree with the higher-scoring range in which current systems now operate (Peyrard, 2019). Based on a summary benchmark SummEval that incorporates in a consistent and comprehensive fashion both metrics and models, we were able to select from 14 metrics a subset that correlates well with human judgments while highlighting areas of opportunity for future work (Fabbri et al., 2021). We build upon this line of research, seeking to understand the relationship between four quality dimension for human judgements (coherence, consistency, fluency, and relevance) and highest correlated metrics. Our main goals are:

- to find out wether 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 toolkit that facilitates the scoring in each of the 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. Human annotated data is expensive and can become obsolete when compared to the state-of-the-art summarization models (Peyrard, 2019). Modern systems provide high quality summaries (Pegasus, T5) while most of the annotated data collected in DUC/TAC[1] conferences tend to generate average summary scores. This is problematic because there's is a lack of high score

[1] http://tac.nist.gov

summaries to evaluate these novel architectures. Abstractive models introduce novel n-grams to a summary, therefore overlapping n-grams are unable to rate a model's quality. Researchers are forced to introduce new metrics. 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 without requiring annotated reference sentences. These are some examples of an estimated twenty three different metrics that have been use to evaluate summaries in the last decade. 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 the Text Analysis Conference (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) and have been used in MT and text generation. However, none of these combined metrics have been used to test summaries, and particularly, moved away from the correlation with the Workshop on Statistical Machine Translation WMT human scores.

3 Methods

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

3.1 Evaluation Metrics

The selection of evaluation metrics includes primarily metrics from the last years directed at MT and summarization based on 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.

3.2 Modeling

In the following we aim to model expert evaluation scores given calculated metrics in a typical discriminative supervised learning setting. Given a dataset $\mathcal{D} = \{(\boldsymbol{X_i}, y_i)\}_{i=1,\dots n}$ 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 $f(x_1, \dots, x_m)$, one would want to quantify to what extent each x_j contributes to the expert evaluation score output. In order to relate the input features 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 settings 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 prediction value among the features 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 unobserved, 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 Data

Here's a subsection within methods

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 X we present key model performance results on the test set.

5.1 Coherence

6 Conclusions

Add Conclusions here

Acknowledgments

The acknowledgments should go immediately before the references. Do not number the acknowledgments section. **Do not include this section when submitting your paper for review.**

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.

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.

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*.

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.

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.

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

Type of Text	Font Size	Style
paper title	15 pt	bold
author names	12 pt	bold
author affiliation	12 pt	

Table 1: NAME OF TABLE

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 Latex tricks

Emphasis in a line: *e.g.*,

quote intentionally a line (**?**) ...

Bold a line: **Papers that do not conform to these requirements may be rejected without review.**

Add bullet points

- Left and right margins: 2.5 cm
- Top margin: 2.5 cm

Paragraph start without indentation: This paragraph is not indented.

Create tables

Table 1 will refer to the table on the text.

Add footnote: They may be numbered or referred to by asterisks or other symbols.[2]

More table formatting:

Hyperlinks colors: Dark Blue text, Color Hex #000099.

Types of citation commands:

[2]This is how a footnote should appear.

Command	Output	Command	Output
{\"a}	ä	{\c c}	ç
{\'u}	ú	{\v r}	ř
{\aa}	å	{\ss}	ß

Table 2: Example commands for accented characters, to be used in, *e.g.*, BIBTEX names.

Output	natbib command	Old ACL-style command
(?)	\citep	\cite
?	\citealp	no equivalent
?	\citet	\newcite
(?)	\citeyearpar	\shortcite

Table 3: Citation commands supported by the style file.