What Makes a Good Podcast Summary?

Conference Paper · July 2022

DOI: 10.1145/3477495.3531802

4 authors, including:

Rezvaneh -Shadi- Rezapour
Drexel University
34 PUBLICATIONS **266** CITATIONS

SEE PROFILE

What Makes a Good Podcast Summary?

Rezvaneh Rezapour
Drexel University
Philadelphia, United States
shadi.rezapour@drexel.edu

Sravana Reddy*
ASAPP
New York, United States
sravana.reddy@gmail.com

Rosie Jones
Spotify
Boston, United States
rjones@spotify.com

Ian Soboroff
NIST
United States
ian.soboroff@nist.gov

ABSTRACT

Abstractive summarization of podcasts is motivated by the growing popularity of podcasts and the needs of their listeners. Podcasting is a markedly different domain from news and other media that are commonly studied in the context of automatic summarization. As such, the qualities of a good podcast summary are yet unknown. Using a collection of podcast summaries produced by different algorithms alongside human judgments of summary quality obtained from the TREC 2020 Podcasts Track, we study the correlations between various automatic evaluation metrics and human judgments, as well as the linguistic aspects of summaries that result in strong evaluations.

CCS CONCEPTS

• **Information systems** → **Summarization**; **Evaluation of retrieval results**; • **Computing methodologies** → *Natural language processing*.

KEYWORDS

podcast summarization, abstractive text summarization, evaluation, ROUGE

ACM Reference Format:
Rezvaneh Rezapour, Sravana Reddy, Rosie Jones, and Ian Soboroff. 2022. What Makes a Good Podcast Summary?. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3477495.3531802

1 INTRODUCTION

A typical podcast episode is about half an hour long, requiring a substantial investment of listening time. Surveys show that users rely on the written text description of the podcast or episode in deciding whether to listen [23]. Podcast summaries could serve as the basis for this decision-making, or even stand in as a synopsis for the full episode when a listener does not have time for the complete listening experience [35]. Automatic summarization could aid podcast creators in writing descriptions, serve to augment manually written descriptions in podcast streaming platforms, or assist in the construction of audio trailers. Podcast summarization is a recently introduced task [3, 12, 13] that raises questions about evaluation. Which features and metrics characterize good summaries for podcasts? Are these the same attributes that predict quality of news article summaries? We study different automatic evaluation metrics and linguistic features, and explore how they correlate with human judgments, in order to quantify what makes a good podcast summary.

2 TREC PODCAST SUMMARIZATION TASK

Our data consists of system submissions for the summarization task of the TREC Podcasts Track [12] and the associated human quality judgments provided by the National Institute of Standards and Technology (NIST)[1].

2.1 Podcast Corpus

The Spotify Podcast Dataset [3, 12] consists of 105,360 podcast episodes, which were designed to be used as training data for the TREC Podcast Summarization Task. Each episode is associated with an automatically generated transcript, the audio of the episode, and its RSS header. The episodes are accompanied by short descriptions of the episodes written by the podcast creators (also referred to as the 'creator's descriptions'). A 'filtered' version of the descriptions was provided in which extraneous content such as boilerplate, ads, promotions, and notes that did not directly describe the episode were removed [31]. An additional 1,027 episodes were released for the task as the test set with the same metadata.

2.2 Summarization Systems

The stated task was to generate a short, accurate, and grammatically sound text summary for each podcast episode using the transcripts of the podcast episodes and/or the original audio. In total, 8 participants submitted 22 models (Table 1) [12]. Systems largely used abstractive techniques, with the BART transformer model [19] trained on news summarization[2] and fine-tuned using the creator's descriptions as targets were the most predominant [14, 22, 33, 34, 38]. Two

*Work was done while at Spotify.

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531802

[1] http://www.nist.gov
[2] https://huggingface.co/facebook/bart-large-cnn

submissions, [26] and U Texas Dallas, used T5 [29] and another one leveraged GANs [8, 15].

A challenge with podcast transcripts is that they tend to be long documents. Due to the maximum input length of BART (1024 tokens; about 10 minutes of audio), some systems used extractive techniques to select the salient sections or sentences from the podcast transcripts. One team [22] used ensemble models, which resulted in the highest human evaluation scores compared to the other submissions, and one team [14] replaced the attention layers of BART with the attention mechanism used in the Longformer [1] to extend the input length to 4096 tokens.

The 'baseline' models, provided by the organizers [12], consisted of off-the-shelf techniques including TextRank, the first one minute of the transcript, and BART, pre-trained and fine-tuned on the training data. None of the submitted models made use of the audio directly. An overview of all systems is provided in Table 1. More details are provided in the Appendix (§A.1).

Participant	Model	method
U New Hampshire	unhtrema1	GAN, LSTM, 3 sentences, long chunks
	unhtrema2	GAN, LSTM, 10 sentences, long chunks
	unhtrema3	GAN, LSTM, 20 sentences, short chunks
	unhtrema4	GAN, LSTM, 10 sentences, short chunks
U Central Florida	UCF_NLP1	BART
	UCF_NLP2	BART, RoBERTa
U Texas Dallas	UTDThesis_Run1	T5, fine tuned on brass set + Dialogue Action Tokens
U Glasgow	2306987O_abs_run1	T5, fine tuned on description
	2306987O_extabs_run2	15 sentence input, T5
	2306987O_extabs_run3	Extractive filtering, SpanBert
U Cambridge	cued_speechUniv1	BART, sentence filtering, 9 model ensemble
	cued_speechUniv2	BART, sentence filtering, 3 model ensemble
	cued_speechUniv3	BART, Fine tuned on transcript
	cued_speechUniv4	BART, sentence filtering, non-ensemble
Uppsala U	hk_uu_podcast1	BART, Longformer, 3 epochs
Spotify	categoryaware1	BART, Fine tuned on start of transcript podcast category; 1 epoch
	categoryaware2	BART, Fine tuned on start of transcript podcast category; 2 epochs
	coarse2fine	BART, Fine tuned on TextRank center of transcript; 2 epochs
U Delaware	udel_wang_zheng1	Start of transcript, BART
	udel_wang_zheng2	Select sentences by LDA, BART
	udel_wang_zheng3	Select sentences by ROUGE, BART
	udel_wang_zheng4	Ensemble of 1-3
Baseline	bartcnn	BART, No fine tuning
	bartpodcasts	BART, Fine tuned on start of transcript
	onemin	1 minute of transcript
	textranksegments	TextRank, 50 wd segments
	textranksentences	TextRank, sentence split

Table 1: Technologies employed by the submitted systems to the TREC Podcast Summarization Task.

2.3 Manual Evaluation

NIST assessors judged each summary on a four-point scale (Excellent, Good, Fair, and Bad) representing how well it conveyed the main gist of the whole podcast episode. The assessments were converted into a numerical EGFB score by a weighting scale of 4-2-1-0, with 4 representing. The detailed evaluation scale is in the Appendix (§A.2).

Furthermore, a set of boolean attributes were assessed; these attributes were derived from a small-scale survey of podcast listeners. These attributes were; *Q1.* presence of names of the main people and characters of the podcast, *Q2.* presence of biographies of the people mentioned, *Q3.* presence of the main topics of the podcast, *Q4.* whether the summary indicated the format and style of the episode, *Q5.* whether the summary provided context to the title of the podcast, *Q6.* whether the summary did not contain redundant information, *Q7.* whether it was written in good English, and *Q8.* whether it started and finished with appropriate and coherent sentences.

The assessors evaluated the summaries of 179 randomly selected episodes produced by all 22 submitted systems, and 5 baseline systems, as well as the 2 human summaries (creator's original and 'filtered' version); giving a set of 5,191 summaries with human judgments. Each summary received one assessment.

To establish the significant differences between the systems as judged by human evaluation, we applied a bootstrap sampling procedure, randomly selecting a sample of 50 summaries, calculating the mean system EGFB score based on the samples, and ranking the systems based on the mean, repeated 100 times. Our results (Fig. 1) show that the differences between the systems are significant. The ranking of systems by EGFB mean scores (Fig. 2) show that the best performing systems were all built upon BART. Systems that reduced redundancy through extractive techniques like sentence or section filtering had high performance in general.

The creator's written descriptions of their podcasts did not score as high as some of the automatically generated summaries, even though creator's descriptions in the training data were used as output targets for training the models and, notably, are used as the ground truth reference for automatic evaluation.[3] This observation suggests that automatic metrics using these descriptions as the reference are necessarily going to be imperfect.

[3]This is likely because of the varying quality of the descriptions across different episodes, and also because descriptions are not necessarily written to serve as summaries.

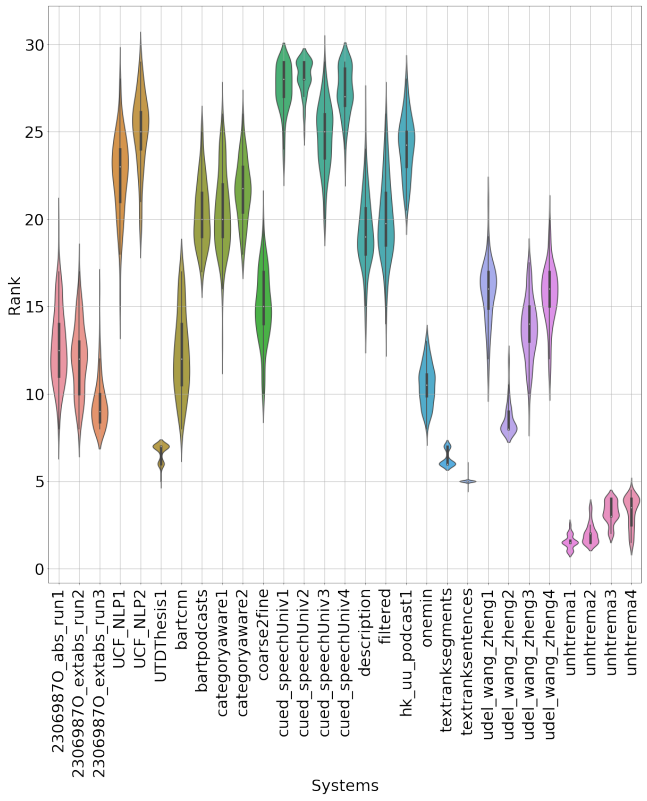

Figure 1: Distribution of system ranks with bootstrapped sampling showing significant differences between systems

3 FEATURES AND EVALUATION METRICS

We selected the following automatic evaluation metrics and features to characterize podcast summaries. Many of the features were computed with the SummEval library [6] and are detailed in the Appendix (§A.3).

Syntactic Features: the proportion of different parts of speech in the summaries.[4]

Readability: length and various readability and complexity statistics (Flesch-Kincaid grade, SMOG index, and Coleman-Liau scores).[5]

Semantic Similarity between the summaries and the podcast transcripts. We computed the cosine similarity between the two using two representations: average word2vec[6] [24], and TF-IDF scores of the words. We also computed measures defined by Grusky et al. [9]: the 'extractiveness' score, and the percentage of n-grams that are 'novel' (n-grams in the summary not present in the transcript) and 'repeated' (n-grams that are present in the transcript).

Reference Comparison Features: where the summary is compared with the 'filtered' creator's written description as the reference, using several evaluation metrics such as **ROUGE** [20], **ROUGE-WE** [25], **BLEU** [27], **METEOR** [18], **CIDEr** [36], **Bert Score** [37], and **chrF** [28].

[4]spaCy [11] was used for tagging.

[5]We computed readability scores with *Textstat*, https://github.com/shivam5992/textstat.

[6]Implementation from Gensim [32].

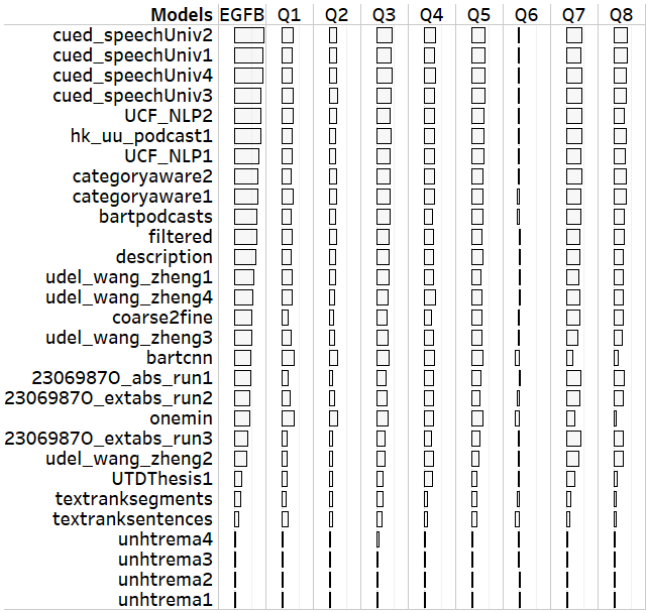

Figure 2: Human evaluation scores of the summaries averaged across the test set for the summarization models (sorted by EGFB). See §A.1 for details.

4 RESULTS

Can automatic evaluation metrics predict human judgment? We used multinomial logistic regression (5-fold cross validation) to investigate how the automatic features and metrics described in Section §3 contribute to the predictability of EGFB scores. We first built a classifier using ROUGE-L[7] metrics as features (i.e., ROUGE-L precision, recall, and f-scores), since ROUGE-L was used for the automatic evaluation of summaries in the TREC Podcast Summarization Task. We then used other metrics (individually or combined) to analyze their impact on the prediction of EGFB score.

As shown in Table 2, BertScore and ROUGE-1 are more predictive of human judgments compared to ROUGE-L. Including the mean EGFB score of each annotator as a feature, representing their bias (a.k.a. 'annotator mean'), into the model gives a considerable boost in predictability. Part of speech features are additionally valuable. A combination of all features results in a considerable improvement in AUC (68.61%) compared to ROUGE-L (59.14%). See Table 4 in the Appendix for more analysis of the features.

While the classifiers are reasonably predictive of human judgments, they may have struggled due to noises in the judgments. As an example, we identified summaries in the set of 5,191 that were word-to-word identical with another and found that 337 of the summaries had duplicates. Within this group, 11.97% were annotated with *different* EGFB scores, although they were generated from the same podcast episodes and were evaluated by the same annotator.

Which features are correlated with good summaries? We computed Kendall's Tau to estimate the marginal correlations between each of the features and the human judgments.

[7]The Longest Common Subsequence

features	Precision	Recall	F1	ROC-AUC
ROUGE-L	41.91	42.56	37.59	59.14
ROUGE-1	43.18	41.98	39.37	60.22
ROUGE-2	40.92	43.69	37.66	58.78
ROUGE-WE-3	41.34	42.37	38.19	58.9
BertScore	46.13	45.95	42.99	62.38
ROUGE-L + annotator mean	44.86	45.87	44.10	61.28
ROUGE-L + annotator mean + length + parts of speech	51.51	49.53	50.10	65.49
all features	**56.16**	**53.73**	**54.52**	**68.61**

Table 2: Prediction of human EGFB scores using a logistic regression classifier with different features and metrics

As shown in Figure 3, various flavors of ROUGE and BertScore are the most correlated with human judgments of quality compared to the other metrics. This is consistent with previous findings on news summarization [6]. The presence of proper nouns ('NNP' and 'NNPS') in summaries appears to be highly correlated as well, as are adverbs ('RBS') and determiners ('DT'). This finding is consistent with the strong correlation between EGFB score and *Q1.* (presence of names of the main people and characters of the podcast). Verbs tend to be inversely correlated with quality ('VBD, VBZ, VBG, VBN, VBP', and 'VB').

A qualitative review indicates that summaries with many proper nouns and determiners tend to be information-dense and specific in contrast to other summaries, particularly those with a preponderance of verbs (Table 3). We also find that extractive density and the percentage of n-grams repeated from the input are correlated with quality, showing that faithfulness is a significant contributor. Summary length ('summ_length') is nearly uncorrelated with human judgments, except for redundancy (*Q6.*). The use of novel n-grams and readability ('summ_flesch-Kincaid, summ_smog, summ_readibility, summ_coleman', and 'summ_dale') are correlated with summaries judged to be non-redundant and written in good English, but not highly correlated with quality.

5 RELATIONSHIP TO PREVIOUS WORK

The question of characterizing or predicting summary quality has been studied in established domains. Bhandari et al. [2] assess the reliability of various metrics in evaluating the output of summarization models using CNN/Daily Mail and TAC 2008+2009 datasets [4, 5] and show that there is no metric that fits all the models and datasets. This finding was one of the motivations for our study. While Kryscinski et al. [16] report that ROUGE only weakly correlates with human assessment, Fabbri et al. [6] benchmark a suite of models for the CNN/Daily Mail corpus [10], and compare them with respect to different evaluation metrics and human judgments. Like us, they find that ROUGE is more correlated with human judgments than other metrics. Louis and Nenkova [21] propose reference-free features of the summary, such as the distributional similarity between the input and the summary. They find that ROUGE is an adequate metric, but that in the absence of reference summaries, semantic similarity features are predictive of human evaluation,

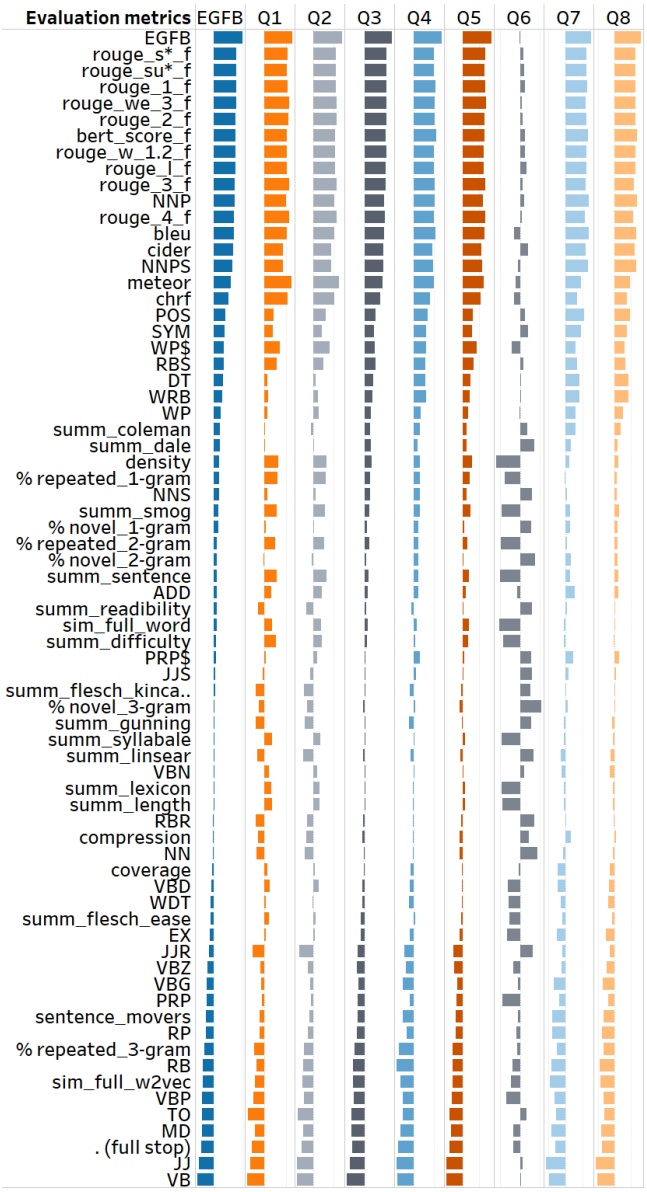

Figure 3: Kendall's Tau correlation between automatically computed features and experts' evaluations of EGFB values and the 8 boolean attributes described in Section §2.3 (converted to numerical judgments).

consistent with our findings. Grusky et al. [9] analyze the extractiveness strategies in a corpus of news summaries with various metrics that we adopted in this paper. Kryscinski et al. [17] design a weakly-supervised model to classify factual consistency of summaries. They show that measuring factuality is a challenging task, and that automated metrics do not correlate well with human judgments. Falke et al. [7] evaluate the factual consistency of summaries with crowdsourcing, and find that ROUGE scores fall short for capturing factual correctness. Since the judgments in our data did not explicitly capture factuality, the problem of predicting

EGFB score	Summary
Excellent	The Nati gang is joined by special guest Tyler Cordy to talk about the Nati Game, Taylor Swift, and more!
Fair	This week, we talk about the upcoming season of The Bachelor. We also discuss some of the biggest names in sports and what they are looking forward to for the rest of the year. Enjoy!

Table 3: Two system-generated summaries for the same podcast episode. The first has a high density of proper nouns that carry specific information, while the second is more generic, with parts of it applicable to many podcasts.

factuality in podcast summaries is still open. While the problem of predicting podcast summarization quality has not been previously studied, there has been work on predicting listener engagement from linguistic signals such as length, part of speech distributions, and others [30].

6 CONCLUSION

We present the first analysis of evaluation for podcast summarization, a domain that is considerably different from well-benchmarked domains like news. We find that high-quality summaries tend to use proper nouns, determiners, and adverbs, and are less likely to use verbs. They also contain more segments repeated from the input. Overall, our results highlight that it is difficult to predict human evaluations even with a suite of several metrics and features. This is partly due to the limitations of automatic metrics (especially given that the reference summaries that are used for metrics like ROUGE are imperfect). We also note that human evaluations are noisy and biased, and that multiple annotations may be valuable.

We believe that the EGFB catch-all score does not capture the nuances of summary quality, and more fine-grained evaluation is needed to assess factors like faithfulness and conciseness. Finally, we find that while ROUGE is not a perfect predictor of quality, it has potential for evaluating podcast summaries and tends to correlate with human judgments better than other metrics. Our results are consistent with previous findings on news summarization.

A APPENDIX

A.1 Summarization Systems

U New Hampshire [15] leveraged Generative Adversarial Networks (GAN) to propose four models (*unhtrema1, unhtrema2, unhtrema3, unhtrema4*) in which the most salient segments with different sizes are selected from the transcripts (using a fixed size blocks) and then summaries are generated using GAN.
U Central Florida [34] proposed two abstractive summarization models (*UCF_NLP1, UCF_NLP2*) using BART and leveraged word saliency to select different segments from the beginning and end of each transcript to generate the summaries.
U Texas Dallas generated summaries using a T5 model fine-tuned on the training data.
U Glasgow [26] used T5 to develop three summarization models (*2306987O_abs_run1, 2306987O_abs_run2, 2306987O_abs_run3*) in which different sentence selection pipelines were tested.
U Cambridge [22] submitted four summarization models (*cued_speechUniv1, cued_speechUniv2, cued_speechUniv3, cued_speechUniv4*) in which redundant sentences were filtered from the input transcripts using the attention of a hierarchical model and a fine-tuned BART and ensembles of three and nine models were used to generate summaries.
Uppsala U [14] proposed a version of the BART summarization model (*hk_uu_podcast1*) by replacing the attention layers with the attention mechanism used in the Longformer to increase the number of input tokens.
Spotify [33] proposed three summarization models (*categoryaware1*, *categoryaware2*, coarse2fine) that take podcasts' genres as well as named entities into consideration in order to generate summaries using fine-tuned BART models.
U Delaware [38] proposed an approach by first extracting the most salient sentences from the transcripts, with respect to the overall information and topics that they covered. They then used BART to develop four models (*udel_wang_zheng1, udel_wang_zheng2, udel_wang_zheng3, udel_wang_zheng4*) to generate the summaries.
Baseline [12] was provided by the organizers of the TREC Podcast Summarization Task. The baselines include a combination of extractive and abstractive models including *onemin*, *bartcnn*, *bartpodcasts*, *textranksentences*, and *textranksegments*.

A.2 Manual Evaluation Scale

Summaries are judged on a four-step scale intended to model how well a listener is able to make a decision whether to listen to a podcast or not, conveying a gist of what the user should expect to hear listening to the podcast. The assessment scale used by the NIST assessors is the EGFB scale, as per the following instructions:
Excellent: the summary accurately conveys all the most important attributes of the episode, which could include topical content, genre, and participants. In addition to giving an accurate representation of the content, it contains almost no redundant material which is not needed when deciding whether to listen. It is also coherent, comprehensible, and has no grammatical errors.
Good: the summary conveys most of the important attributes and gives the reader a reasonable sense of what the episode contains with little redundant material which is not needed when deciding whether to listen. Occasional grammatical or coherence errors are acceptable.
Fair: the summary conveys some attributes of the content but gives the reader an imperfect or incomplete sense of what the episode contains. It may contain redundant material which is not needed when deciding whether to listen and may contain repetitions or broken sentences.
Bad: the summary does not convey any of the most important content items of the episode or gives the reader an incorrect or incomprehensible sense of what the episode contains. It may contain a large amount of redundant information that is not needed.

A.3 Evaluation Metrics

ROUGE [20] compares the produced summaries against the transcripts by measuring the overlap between the n-grams or any other sequences of words in their texts. Depending on the textual unit that is used to calculate recall, ROUGE can be in the form of ROUGE-N

(n-gram), ROUGE-L (the longest common subsequence), ROUGE-W (weighted longest common subsequence), and ROUGE-S (skip-bigram).

ROUGE-WE [25], enhances ROUGE by using word2vec embeddings [24] to calculate the cosine similarity of the summaries and references.

BLEU [27] measures the overlap between matching n-grams in the summaries and reference. This metric leverages a modified n-gram precision to account for brevity.

METEOR [18] measures the harmonic mean of unigram precision and recall, accounting for stemming, synonyms, and paraphrase matching.

CIDEr [36] calculates the co-occurrences of n-grams (n = [1-4]) and the cosine similarity between them in the reference sentences and summaries after stemming.

BertScore [37], computes the similarity scores between the references and summaries by aligning them on a token-level and using embeddings from BERT.

chrF [28], calculates the n-gram character overlap between the references and generated summaries.

Semantic Similarity Statistics as implemented by Fabbri et al. [6] and Grusky et al. [9] compute the 'extractiveness' of a summary with three different statistics: *coverage;* the percentage of words in the summary that are from the input, *density;* the average length of the fragment to which each word in the summary belongs, and *compression ratio;* the ratio of the length between the input and its summary. They also compute the percentage of n-grams in the summaries that are not present in the input, and the percentage of n-grams in the summary that are present.

features	Precision	Recall	F1	ROC-AUC
ROUGE-L	41.91	42.56	37.59	59.14
NETEOR + BLEU + CIDEr	40.08	42.04	37.73	58.29
BertScore	46.13	45.95	42.99	62.38
CHRF + data statistics	42.82	41.23	40.55	59.62
ROUGE-1	43.18	41.98	39.37	60.22
ROUGE-2	40.92	43.69	37.66	58.78
ROUGE-4	35.6	41.94	31.04	54.28
ROUGE-W-1.2	40.18	40.9	35.09	57.48
ROUGE-S*	41.62	43.38	38.61	59.3
ROUGE-SU*	42.31	43.59	38.86	59.57
ROUGE-WE-3	41.34	42.37	38.19	58.9
ROUGE-L + annotator mean	44.86	45.87	44.10	61.28
ROUGE-L+ annotator mean + length	44.49	45.43	43.82	61.08
ROUGE-L + annotator mean + length + parts of speech	51.51	49.53	50.10	65.49
ROUGE-L + length + parts of speech + semantic sim + annotator mean	51.74	49.82	50.38	65.67
ROUGE-L + length + parts of speech + semantic sim + readability + annotator mean	54.39	51.52	52.48	67.13
all features combined	**56.16**	**53.73**	**54.52**	**68.61**

Table 4: Comparative predictability of EGFB scores with different feature combinations

A.4 Correlations between Metrics and Human Evaluation

Table 4 shows a detailed regression analysis of how the automatic metrics and features predict human judgments. Figure 4 shows all the pairwise correlations between features.

REFERENCES

[1] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).

[2] Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating Evaluation in Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 9347–9359. https://doi.org/10.18653/v1/2020.emnlp-main.751

[3] Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. 100,000 Podcasts: A Spoken English Document Corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 5903–5917. https://doi.org/10.18653/v1/2020.coling-main.519

[4] Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the TAC 2008 Update Summarization Task.. In *TAC*.

[5] Hoa Trang Dang and Karolina Owczarzak. 2009. Overview of the TAC 2009 summarization track. In *proceedings of the Text Analysis Conference*.

[6] Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. arXiv:2007.12626 [cs.CL]

[7] Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2214–2220. https://doi.org/10.18653/v1/P19-1213

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[9] Max Grusky, Mor Naaman, and Yoav Artzi. 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)*. Association for Computational Linguistics, New Orleans, Louisiana, 708–719. https://doi.org/10.18653/v1/N18-1065

[10] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems (NIPS)*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates. http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.pdf

[11] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-strength Natural Language Processing in Python*. https://doi.org/10.5281/zenodo.1212303

[12] Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth JF Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu. 2020. TREC 2020 Podcasts Track Overview. In *The 29th Text Retrieval Conference (TREC) notebook. NIST*.

[13] Rosie Jones, Hamed Zamani, Markus Schedl, Ching-Wei Chen, Sravana Reddy, Ann Clifton, Jussi Karlgren, Helia Hashemi, Aasish Pappu, Zahra Nazari, et al. 2021. Current Challenges and Future Directions in Podcast Information Access. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[14] Hannes Karlbom and Ann Clifton. 2020. Abstractive Podcast Summarization using BART with Longformer attention. In *The 29th Text Retrieval Conference (TREC) notebook. NIST*.

[15] Sumanta Kashyapi and Laura Dietz. 2020. TREMA-UNH at TREC 2020. In *The 29th Text Retrieval Conference (TREC) notebook. NIST*.

[16] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural Text Summarization: A Critical Evaluation. In

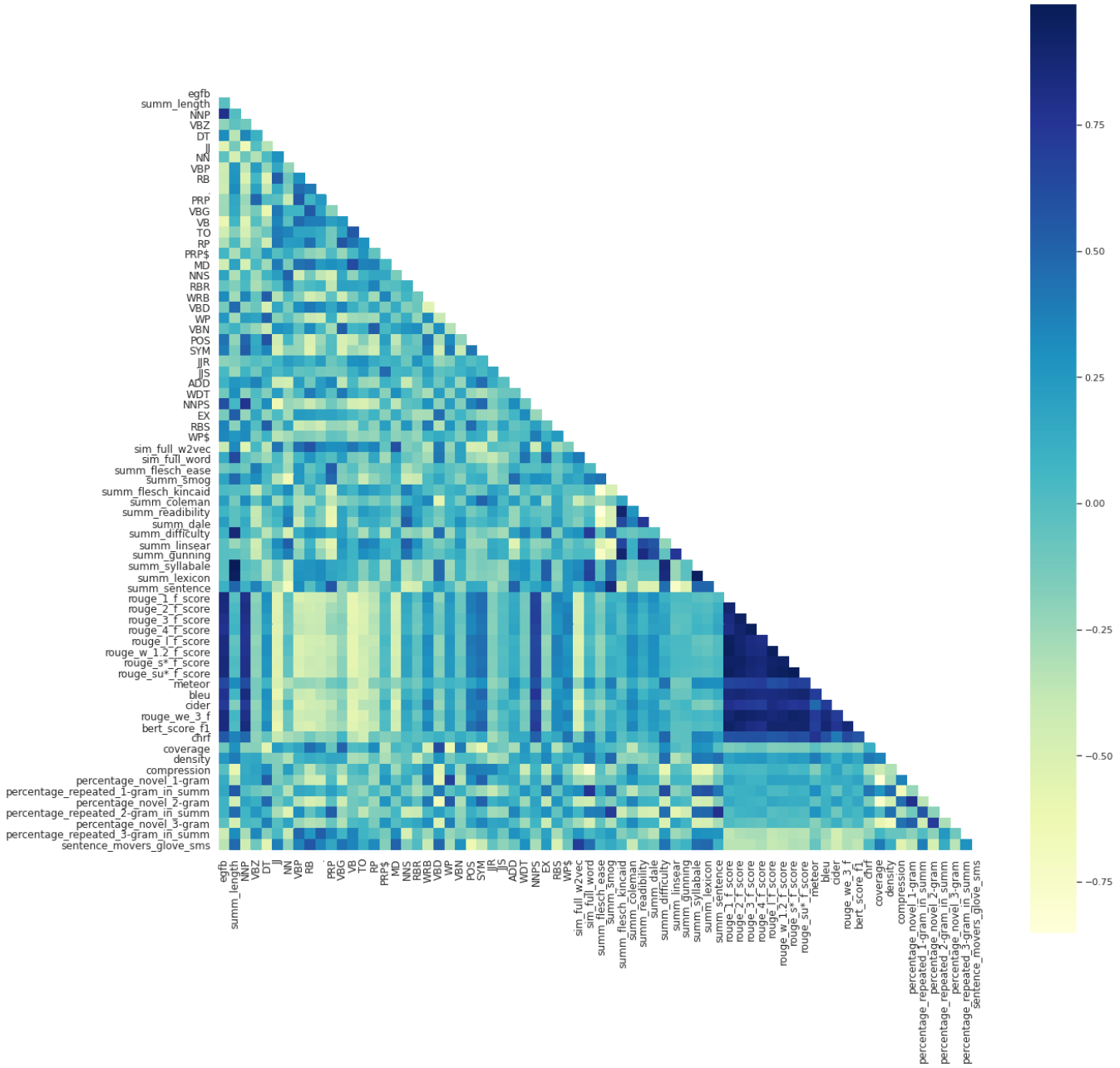

Figure 4: Pairwise Kendall's Tau correlations for all automatically computed features.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 540–551. https://doi.org/10.18653/v1/D19-1051

[17] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 9332–9346. https://doi.org/10.18653/v1/2020.emnlp-main.750

[18] Alon Lavie and Abhaya Agarwal. 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*. Association for Computational Linguistics, Prague, Czech Republic, 228–231. https://aclanthology.org/W07-0734

[19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

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

[21] Annie Louis and Ani Nenkova. 2009. Automatically Evaluating Content Selection in Summarization without Human Models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 306–314. https://aclanthology.org/D09-1032

[22] Potsawee Manakul and Mark Gales. 2020. CUED_speech at TREC 2020 Podcast Summarisation Track. In *The 29th Text Retrieval Conference (TREC) notebook. NIST*.

[23] Matthew McLean. 2020. Podcast Discovery Stats in 2020: How Listeners Discover New Shows. *The Podcast Host* (Dec 2020). 'https://www.thepodcasthost.com/promotion/podcast-discoverability/' Accessed Dec 2020.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[25] Jun-Ping Ng and Viktoria Abrecht. 2015. Better Summarization Evaluation with Word Embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 1925–1930. https://doi.org/10.18653/v1/D15-1222

[26] Paul Owoicho and Jeff Dalton. 2020. Glasgow Representation and Information Learning Lab (GRILL) at TREC 2020 Podcasts Track. In *The 29th Text Retrieval Conference (TREC) notebook. NIST*.

[27] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[28] Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, 392–395. https://doi.org/10.18653/v1/W15-3049

[29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[30] Sravana Reddy, Mariya Lazarova, Yongze Yu, and Rosie Jones. 2021. Modeling Language Usage and Listener Engagement in Podcasts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 632–643. https://doi.org/10.18653/v1/2021.acl-long.52

[31] Sravana Reddy, Yongze Yu, Aasish Pappu, Aswin Sivaraman, Rezvaneh Rezapour, and Rosie Jones. 2021. Detecting Extraneous Content in Podcasts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*.

[32] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50.

[33] Rezvaneh Rezapour, Sravana Reddy, Ann Clifton, and Rosie Jones. 2021. Spotify at TREC 2020: Genre-Aware Abstractive Podcast Summarization. In *The 29th Text Retrieval Conference (TREC) notebook. NIST*.

[34] Kaiqiang Song, Chen Li, Xiaoyang Wang, Dong Yu, and Fei Liu. 2020. Automatic summarization of open-domain podcast episodes. In *The 29th Text Retrieval Conference (TREC) notebook. NIST*.

[35] Damiano Spina, Johanne R Trippas, Lawrence Cavedon, and Mark Sanderson. 2017. Extracting audio summaries to support effective spoken document search. *Journal of the Association for Information Science and Technology* 68, 9 (2017), 2101–2115.

[36] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.

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

[38] Chujie Zheng, Kunpeng Zhang, Harry Jiannan Wang, and Ling Fan. 2020. A Two-Phase Approach for Abstractive Podcast Summarization. In *The 29th Text Retrieval Conference (TREC) notebook. NIST*.