Performance of Summary Machine Evaluators Across Different Model Types

Alex Liu
McGill University
Montreal, Canada
alex.liu
@mail.mcgill.ca

Filippo Baronello
McGill University
Montreal, Canada
filippo.baronello
@mail.mcgill.ca

Ningchen Ma
McGill University
Montreal, Canada
ningchen.ma
@mail.mcgill.ca

Abstract

We don't always have the time or the will to read things written for us. Sometimes, we would simply like to know if we even should read something in the first place. Enter summaries. They are on the back of every book, at the header of every article and in the abstract of every paper. They are fundamental to the efficient extraction of key contents by human readers. As such, recent work in NLP has undertaken the task of generating these summaries through machine learning models. However, the problem remains of having a machine recognize a good abstract. Multiple automated evaluators exist in that regard. But there is an issue with how evaluation metrics are evaluated themselves and whether an evaluator actually measures what it claims to. In this paper, we make the assumption that the gold standard of summary evaluation is human evaluation. Summaries are, after all, written for humans. As such, we conduct blind human evaluations on the output of 4 summary models (2 extractive, 2 abstractive) ran on the CNN/Dailymail dataset(Nallapati et al., 2016). Then, we measure their correlation with several machine evaluators to see if their performance varies across different types of models. We found that the Rouge metrics, while harsher on abstractive models, were better correlated to human evaluation on extractive models. CHRF performance was noticeably better for abstractive models.

Code can be found: here

1 Introduction

1.1 Overview

With the assumption that human evaluations are the gold standard for summary model evaluation, comes the obvious reality that humans are slow. There is simply no efficient way for humans to pour through the corpora of summary generation models every time someone wants to train one. However, when it comes to machine evaluators (we will simply call them metrics from now on), a new problem arises consisting of ascertaining their quality. In essence, how do we know if a metric can faithfully represent the quality of a model. In this paper, we first choose to make the assumption that there is no better evaluator than humans. Summaries are written for us, and thus, they shall be judged by us. As such, we define a relatively good metric to be a metric which is highly correlated to human evaluations. I.e. we want a metric that is most representative of human evaluations. We then assume that this, by extension, means that it is most representative of actual summary quality. However, our hypothesis is that that while a metric may be great for judging one aspect of one type of model, it may not be so good for other aspects and/or other models (no free lunch).

As such, we aim to test this by first generating our (all three authors) own personal human evaluations on the model-generated summaries for 50 randomly selected articles. The models we use are PNBERT(Zhong et al., 2019), MatchSum(Zhong et al., 2020), LoopSum(Laban et al., 2020) and PGen(See et al., 2017) (two abstractive, two extractive). By using a script, we then conduct human evaluations for each type of model by blindly rating every pair of summaries from the same article, generated by the same type of model, against each other. We choose to specifically look at two general aspects of the summaries: relevance and coherence. We define relevance as concisely capturing the main points of the input article. We define coherence as the general quality of the output text (syntax, flow, grammar, etc.). As such, for each summary, we then give two scores from 1.0-5.0, one for each of these aspects. Then, we run our metrics (ROUGE-1 to 4, ROUGE-L (Lin, 2004b), METEOR(Lavie and Agarwal, 2007), chrF++(Popović, 2015), BLEU(Papineni et al., 2002)) on the same summaries and look for correlation with the human scores of each aspect.

1.2 Language models used

Briefly, the difference between an extractive and abstractive summarization model is that the former attempts to generate a summary using excerpts taken directly from the original text, while the latter attemps to summarize by paraphrasing using novel sentences. While both of these types have value, it can be surmised that the task of determining which metrics to use might be different between the two types.

As such, PNBERT(Zhong et al., 2019) and MatchSum(Zhong et al., 2020) are both extractive models. Though both are of the same type and seem to be evenly matched, they are quite a bit different in the way they function. As the name suggests, the former is a BERT-based model while the second operates instead based on semantic matching.

As for our abstractive models, PGen(See et al., 2017) and LoopSum(Laban et al., 2020), the story are quite a bit different. For reasons we will explain later, not only do these two models function differently, but they also have significantly different performance levels. Indeed, while the former seems rather evenly matched with its extractive counterparts, the latter fails to reach such heights.

1.3 Evaluation models used

As mentioned above, the metrics we use are ROUGE-1 to 4, ROUGE-L (Lin, 2004b), METEOR(Lavie and Agarwal, 2007), chrF++(Popović, 2015), BLEU(Papineni et al., 2002). These are all metrics which work by comparing an output summary with a reference summary associated with the summarized article. This reference summary is essentially one that we assume to be the gold-standard of summaries for that paper. Concretely, in the frame of reference of an experiment making use of the above metrics to evaluate its model, the ultimate goal of the model is then to output summaries of the same quality as the corresponsing references.

2 Related works

Many previous works exist on examining the correlation of metrics with human scoring of certain summary aspects. However, none that we know specifically examines the performance of metrics from model to model.

For example, a work by Fabbri et al (Alexander R. Fabbri, 2021) examines both the quality of 23 models, as well as the correlation of 13+ metrics with human scoring on those models. It offers little discussion but, instead, focuses on presenting the correlation of metric scores with human scoring of different aspects across all models. That is, for a given metric and human graded aspect (i.e. relevance), a single correlation coefficient is computed which encompasses every score the metric gave across all models. This contrasts with our paper, as we keep the correlation scores not only segregated by aspect, but also by model. This allows us to examine the variation in performance of metrics from model type to model type.

Another work by Lin (Lin, 2004a) specifically examines the effectiveness of variations of ROUGE, a general group of metrics we use in this paper. In it, she measures the Pearson Correlation coefficients of ROUGE scores with respect to human grading on summaries of the DUC dataset. She also examines different techniques related to the reference (gold standard) corpus which ROUGE grades summaries against. This includes the usage of many references for a single article, the removal of stop words, stemming, etc. Once again, there is little examination of performance across models, and correlation is computed with respect to all summaries. It should be noted that a huge number of works exist specifically in the vein of examining ROUGE. For example, another work (Graham, 2015) compares ROUGE with BLEU (another of the metrics we use).

A work by Owczarzak et. al (Owczarzak et al., 2012) shifts instead the scope to summary human evaluation itself. It finds that a lot of human annotation suffers from inconsistency. However, this does not seem to hurt system-level evaluation. That is, for example, the average evaluation of a model's output seems to stay in line with the average metric score. However, on a summary to summary basis, the paper indeed finds a necessity to minimize human inconsistency. Indeed, the paper made use of methods to remove inconsistently scored output summaries and found that the correlation of metrics with the remaining scores increased considerably. While we do not specifically examine human scoring, our whole experiment hinges on it. This paper then outlines a potential limitation to ours. As such, we have taken every step within reason to keep our scores consistent, through the use of carefully defined aspects, as well as blind testing.

3 Method

3.1 Summaries

In this paper, we use 4 pretrained, fairly recent models - two extractive (PNBERT, MatchSum), two abstractive (PGen and LoopSum). All of them were optimized for the CNN/Dailymail dataset(Nallapati et al., 2016). Using existing literature on the models, we tried our best to pick the first 3 of these as evenly matched as possible, in order to provide a level playing field for our metrics. In essence, the idea is to allow the variation between the scores of a metric on different models to reflect, as much as possible, the performance of the metric with respect to the models, rather than the quality of the models. However, we also decided to include Loopsum, which is inferior to the other three, in order to examine how different metrics deal with a gap in quality between models of different types.

3.2 Blind human evaluation

First we want to conduct a blind human evaluation of 200 model-generated summaries of the 50 randomly selected articles from the CNN/Dailymail corpus(Nallapati et al., 2016). As such, this is 4 summaries, generated by 4 different models (PNBERT, MatchSum, LoopSum, PGen) for each of the 50 articles. Then we write a script which, for each article, presents an article and the four of its summaries generated by each model, in random order, with their corresponding models hidden. Each author, independently from each other, is then made to input two floating point scores for each summary - one for relevance and one for coherence, based on their own judgement of the summaries with respect to the chosen aspects. Both of these scores are on a scale from 1.0-5.0. Finally, a csv file with the scores of each model is saved for each aspect.

3.3 Metric evaluations and correlation

We then run a series of metrics on the same 200 summaries. We used off-the-shelf implementations of the following metrics: ROUGE-1 to ROUGE-4, ROUGE-L, METEOR, chrF++ and BLEU. These are a mix of popular metrics and metrics that we believe could succeed where the other ones fail. They all grade against gold standard reference summaries I.e. summaries that are assumed to be near-perfect. In this case, we used the reference summaries provided for that purpose along with the CNN/Dailymail dataset(Nallapati et al., 2016). They appear to be a mix of human written and machine written summaries that were used for advertising purposes. We assume that they are indeed the gold standard.

Finally, for each model, we compute the correlation of each metric with the human scores. I.e. for a given model, we obtain two correlation scores (one for relevance, one for coherence) for every metric used on it. Essentially, we are trying to see how each metric's scores correlate with the human evaluation of the two aspects for a given model.

4 Results

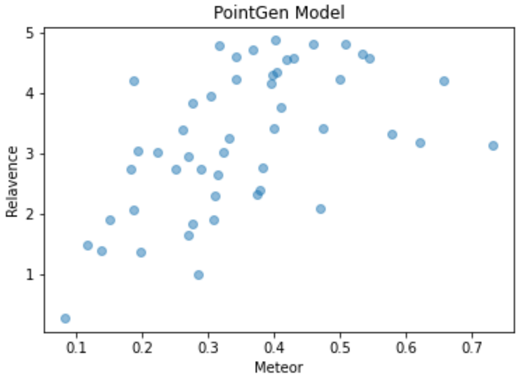

Figure 1: PGEN METEOR scores with respect to human evaluation on relevance. Fairly correlated.

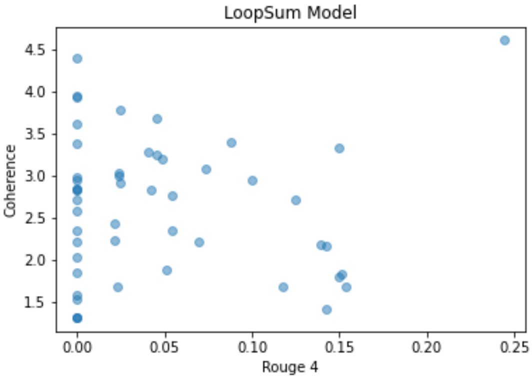

Figure 2: LoopSum ROUGE-4 scores with respect to human evaluation on coherance. Not at all correlated

Full Kendall's Tau correlation results are included in the appendix (Tables 1-4), along with the mean human and metric scores (Table 5). As we can see above. some metrics were decently correlated with human scores of some model (Figure 1.), while others downright weren't (Figure 2.).

5 Discussion and conclusion

5.1 Findings

From our results, as expect, we find that no single metric is great across all models and all aspects of the summaries. First we notice that aggregate human evaluation scores do support our inital belief that MatchSum, PNBERT and PGen are quite evenly matched while LoopSum is a league under. As for the metrics, we notice that most evaluators tended to produce scores related more closely to relevance than coherence.

BLEU seemed to be a top performer across all models with the exception of LoopSum, where its score had virtually no correlation with the human coherance grading. Much of the same is true for the ROUGE metrics, espeically the higher N ROUGE-Ns. They all had decent performance across all models, except for LoopSum. A possible explanation is that, LoopSum, being a rather poor abstractive model that humans and machine graded poorly, frequently spits out an incoherent string of jumbled key statements. As such, this could severely punish metrics that focus on n-grams. Indeed, if we look at the actual results, ROUGE-1 scored LoopSum disproportionately high relative to the human scores, while higher N ROUGE-Ns graded LoopSum disproportionately low. This further adds to our belief that metrics which simply look at the n-grams have trouble accounting for incoherent outputs pieced together with key statements. If relevance is all we're looking for, on the other hand, then BLEU is the clear winner as it performed the best on all models with the exception of PNBERT, where it scored second place by only a little. Otherwise, these metrics might not be suitable for lower end abstractive models.

For chrF++, we found that it performs extremely poorly on extractive models. Indeed, it has almost 0 correlation with human scoring on both relevance and coherance. However, when it comes to abstractive models, it is most consistent with human evaluation as far as a consistent balance of relevance and coherance is required. Indeed, we can see that although it gets outperformed by BLEU and METEOR for PGEN, it is the only one still correlated with humans in terms of coherence for LoopSum, by quite a wide margin. This suggest that chrF++ is the only metric out of the ones we tested that is capable of accounting for coherence in abstractive model summaries.

Finally, when it comes to METEOR, it is by far the most inconsistent one. It performed quite well on the first extractive model, MatchSum while completely failing to account for coherence on the second one, PNBERT. It suffered from the same problems as the ROUGE/BLEU metrics with regards to LoopSum, but, when it comes to PGen, it gave the highest correlation scores of the whole experiment for both relevance and coherence. Due to its inconsistency, we cannot conclude that it's a good metric for high performing abstractive models off of one case. However, due to the impressive improvement in performance it showed for PGen, more experimentation in that regard could prove fruitful.

5.2 Limitations

First of all, the blind aspect of our human evaluation was slightly hurt by LoopSum which consistently output extremely recognizable jumbles of key statements. Furthermore, only having 3 humans evaluators for the summaries of only 4 models on 50 articles leaves room for a lot of bias. We believe that our results were still statistically significant but we agree that more evaluators on more summaries and more models would've greatly improved the significance of our experiment. Unfortunately, the time we had simply didn't allow for more.

When it comes to metrics, we could not implement some of the more complex ones such as the model-based BERTScore due to computer limitations. This could've potentially yielded more interesting results.

5.3 Future work

This work is easily extendable. The first step would likely be to include more models that are bigger and more recent. More metrics would also be in order. Then, as mentioned above, more human testers on more articles would also be in order. This would, by far, be the most difficult improvement to make as grading these summaries requires reading articles of varying length and cross examining it against the presented summaries. I.e. it's quite a lengthy process.

6 Statement of contribution

All team members gave their scores on coherence and relevance for all of the 200 summaries involved in this experiment. They all contributed to the overall design of the experiment as well. This includes defining the hypothesis, outlining the nature of the desired results as well as the processes that should lead to them (blind human evaluations, correlation computation, which models to pick, which metrics to pic).

Alex Liu wrote the paper and implemented the ROUGE metrics.

Filippo Baronello wrote the script to allow for blind human evaluation and computed the final correlation results.

Ningchen Ma generated graphics for the paper, implemented Bleu, Chrf and Meteor evaluators and did the citations.

References

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

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal. Association for Computational Linguistics.

Philippe Laban, Andrew Hsi, John Canny, and Marti Hearst. 2020. The summary loop: Learning to write abstractive summaries without examples. pages 5135–5150.

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*, StatMT '07, page 228–231, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004a. Looking for a few good metrics: Automatic summarization evaluation - how many samples are enough? In *NTCIR 2004 (NTCIR-4)*. NII, NII.

Chin-Yew Lin. 2004b. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Karolina Owczarzak, Peter A. Rankel, Hoa Trang Dang, and John M. Conroy. 2012. Assessing the effect of inconsistent assessors on summarization evaluation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 359–362, Jeju Island, Korea. Association for Computational Linguistics.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

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

Abigail See, Peter Liu, and Christopher Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Association for Computational Linguistics*.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching.

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2019. Searching for effective neural extractive summarization: What works and what's next. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1049–1058, Florence, Italy. Association for Computational Linguistics.

A Appendix

	rouge 1	rouge2	rouge3	rouge4	rougeL	bleu	chrf	meteor
relevance	0.061	0.179	0.203	0.191	0.186	0.269	0.213	0.176
coherence	-0.063	-0.014	0.021	-0.022	-0.002	0.094	0.237	0.093

Table 1: LoopSum correlation between human scores and metrics

	rouge 1	rouge2	rouge3	rouge4	rougeL	bleu	chrf	meteor
relevance	0.232	0.241	0.248	0.243	0.222	0.272	-0.043	0.208
coherence	0.149	0.217	0.270	0.251	0.165	0.253	-0.099	0.227

Table 2: MatchSum correlation between human scores and metrics

	rouge 1	rouge2	rouge3	rouge4	rougeL	bleu	chrf	meteor
relevance	0.324	0.291	0.278	0.266	0.270	0.300	0.092	0.259
coherence	0.210	0.141	0.154	0.181	0.161	0.202	-0.010	0.065

Table 3: PNBert correlation between human scores and metrics

	rouge 1	rouge2	rouge3	rouge4	rougeL	bleu	chrf	meteor
relevance	0.201	0.231	0.232	0.233	0.149	0.308	0.279	0.423
coherence	0.181	0.175	0.154	0.155	0.195	0.207	0.158	0.295

Table 4: PGen correlation between human scores and metrics

	human-relevance	human-coherence	rouge 1	rouge2	rouge3	rouge4	rougeL	bleu	chrf	meteor
MatchSum	3.812	3.534	0.425	0.203	0.124	0.083	0.273	15.629	31.191	0.395
PNSum	3.821	3.312	0.399	0.195	0.115	0.077	0.262	14.934	30.495	0.430
PointGen	3.229	3.374	0.419	0.206	0.128	0.090	0.301	14.709	31.268	0.351
LoopSum	2.943	2.642	0.453	0.188	0.094	0.048	0.287	5.848	24.787	0.222

Table 5: Human and metric score means per model