Evaluation of Review Summaries via Question-Answering

Nannan Huang
RMIT University, Australia
s3754491@student.rmit.edu.au

Xiuzhen Zhang
RMIT University, Australia
xiuzhen.zhang@rmit.edu.au

Abstract

Summarisation of reviews aims at compressing opinions expressed in multiple review documents into a concise form while still covering the key opinions. Despite the advancement in summarisation models, evaluation metrics for opinionated text summaries lag behind and still rely on lexical-matching metrics such as ROUGE. In this paper, we propose using the question-answering(QA) approach to evaluate summaries of opinions in reviews. We propose to identify opinion-bearing text spans in the reference summary to generate QA pairs so as to capture salient opinions. A QA model is then employed to probe the candidate summary to evaluate information overlap between the candidate and reference summaries. We show that our metric RunQA, Review Summary Evaluation via Question Answering, correlates well with human judgments in terms of coverage and focus of information.

1 Introduction

Opinion summarisation takes input documents like online reviews or social media posts where users express their opinions on topics and condense them into a single piece of text. The summary should reflect the core opinions expressed in the source documents. Recent studies in opinion summarisation have shown an advancement moving from extractive (Meng et al., 2012; Ku et al., 2006; *i.a*), i.e. copying sections from the original reviews to produce a summary, to abstractive (Chu and Liu, 2019; Bražinskas et al., 2020a; Bražinskas et al., 2020b; *i.a*), i.e. generating new phrases that reflect the information covered in the original text.

Despite the advancement in summarisation models, evaluation metrics for opinionated text summaries lag behind. For evaluation of review summaries, traditional token-matching ROUGE (Lin, 2004) is still widely used, supplemented with human evaluation on the relative ranking of system-generated summaries in terms of quality dimensions such as Fluency, Coherence, Non-redundancy, Informativeness, and Sentiment (Chu and Liu, 2019; Bražinskas et al., 2020a; Bražinskas et al., 2020b). It is well understood that ROUGE cannot capture the same meaning expressed in different token sequences.

Neural model-based metrics have been proposed for general text summarisation evaluation. Especially QAEval (Deutsch et al., 2021) has been proposed for evaluating the information quality of abstractive summaries with respect to the reference summaries. A key step for the success of QAEval for summarisation evaluation is extracting answers and generating questions covering a significant amount of important Summarisation Content Units (SCUs). Generally, noun phrases(NP) and named entities(NER) are used to generate question-answer pairs in QA models (Durmus et al., 2020; Wang et al., 2020; Deutsch et al., 2021), but their applicability for review summary evaluation is yet to be examined.

In this paper, we propose to evaluate review summaries with question answering (QA) based on neural models, with a focus on evaluating the information quality of review summaries. Modern abstractive summarisation systems can generate sentences of high linguistic quality – grammatically correct, easy to read and understand– but it is more important to evaluate the *information quality* of summaries. Specifically system summaries of high quality information should express opinions consistent with those in the reference summary.

We propose to evaluate the information quality of review summaries, in terms of coverage(recall) and focus(precision) (Koto et al., 2020), where coverage is the amount (proportion) of salient information of the reference summary that the system summary contains, and focus is the amount (propor-

tion) of salient information in the system summary. To improve the QA framework for more effective review summary evaluation, we propose to identify opinion-bearing text spans to generate QA pairs rather than relying on NPs and NERs. In addition to the evaluation of the information quality of summaries, we further propose evaluating the robustness of evaluation metrics for ranking summaries through an adversarial task.

Our evaluation metric RunQA, namely Review Summaries Evaluation via Question Answering, was evaluated against QAEval and other metrics on an Amazon review summarisation dataset (Bražinskas et al., 2020b) We found that RunQA significantly outperforms QAEval and other metrics for evaluating the information quality of summaries, especially in terms of precision. We also found that RunQA is the most robust for ranking summaries.

2 Related Work

We first discuss automatic metrics for general text summarisation evaluation and then especially discuss the QA-based metrics.

2.1 Evaluation Metrics for Text Summarisation

Automatic metrics for summarisation evaluation can be broadly divided into three groups – traditional token matching-based metrics, embedding-based metrics and model-based metrics.

Token matching-based metrics: When evaluating the performance of a summarisation system, researchers introduced metrics by comparing n-gram token matching such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) over a decade ago. Owing to simplicity and ease of use, ROUGE is one of the most widely used automatic metrics. It is designed to capture the similarity between text sequences based on lexical overlaps.

Embedding-based metrics: The significant improvement in the summarisation domain moving from extractive to abstractive makes using lexical overlap metrics for evaluation inadequate. In abstractive summarisation, the summary does not necessarily use the exact word when contextual embeddings like BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018) are used. ROUGE becomes less suitable in this setting, as it performs a surface-level comparison between texts, and it fails to compare words that express the same meaning expressed in different forms.

To overcome the problem of exact word matching, researchers introduced metrics using contextual embeddings, such as BERTScore (Zhang et al., 2019) and MoverScore (Zhao et al., 2019). Both of these metrics have proven to better correlate with human judgments than ROUGE.

Although the embedding-based metrics overcome problems with exact word matching, they are still comparing two pieces of text by capturing context similarity, but not evaluating whether they express the same information (Deutsch and Roth, 2020).

ROUGE is still the most widely used and the default metric for evaluating opinion summarisation. In recent opinion summarisation papers (Chu and Liu, 2019; Bražinskas et al., 2020a; Bražinskas et al., 2020b), the ROUGE family is still the only automatic metric used for evaluating their systems.

Researchers have shown ROUGE is weakly correlated with human judgments (Novikova et al., 2017), and it is not suitable to be used for opinion summarisation evaluation (Tay et al., 2019) or abstractive summarisation (Ng and Abrecht, 2015). It is calculated based on token overlap rather than looking at whether summaries express the same opinion. This makes ROUGE not an ideal metric for evaluating opinionated text summaries.

Both the token overlap-based and embedding-based metrics have the drawback of weakly penalising information or opinion inconsistency (Tay, 2019). For example, for documents expressing opposite opinions like 'I like sushi' and 'I hate sushi'. ROUGE will penalise it weakly by putting equal weight on each token. Whereas embedding-based metrics will treat 'like' and 'hate' similarly because of the similar context.

Model-based metrics: Recent studies have introduced different model-based metrics. For example, SUPERT (Gao et al., 2020) and LS-Score (Wu et al., 2020) target to evaluate text summarisation without references. SUPERT achieved this by generating pseudo references using the top sentences in the source documents, and LS-Score by generating different negative samples and applying unsupervised contrastive learning to learn the metric.

Model-based metrics include QA-based metrics. Following that, we go into metrics based on QA models in further depth.

2.2 QA-based Metrics

There are two types of QA-based metrics targeting evaluation in different dimensions. One type is reference-free models such as FEQA (Durmus et al., 2020) and QAGS (Wang et al., 2020), where questions are generated using the candidate summary, and asked against the source document to measure the faithfulness of the candidate summary. The other type is reference-based models such as QAEval (Deutsch et al., 2021), where questions are generated using a reference summary and asked against the candidate summary to evaluate content and information overlap between the summaries.

QAEval is proven to generate question-answer pairs that cover a significant amount of information expressed in summaries. It also correlates well with human judgments when used as a reference-based metric. Our work builds on QAEval for opinion summarisation evaluation. The original QAEval model generates questions by extracting noun phrases only. In this work, we use a different answer selection strategy to capture and evaluate the information and opinions expressed in summaries.

3 RunQA: Review Summary Evaluation via Question Answering

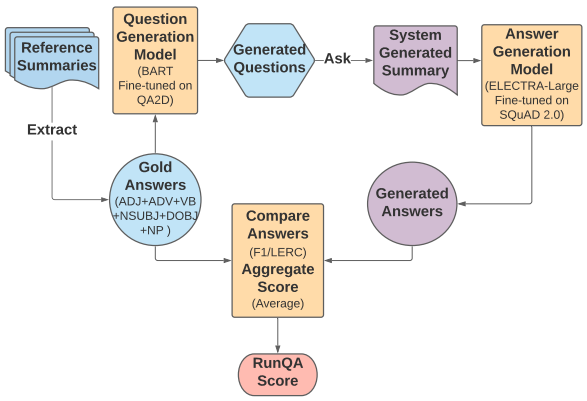

Figure 1: The RunQA model architecture. The answer selection strategy changed from noun phrases(NP) only to the combination of adjectives, adverbs, verbs with their subject child and object child and NP.

The overall architecture of the RunQA model is shown in Figure 1. Like the original model, it consists of a question generation model using a pre-trained BART language model (Lewis et al., 2020) and fine-tuned on QA data provided by Demszky et al. (2018). QAEval uses a pre-trained ELECTRA-Large language model (Clark et al., 2020) and fine-tuned it on the SQuAD 2.0 (Rajpurkar et al., 2018) dataset for question answering.

RunQA has several key differences from QAEval (Deutsch et al., 2021). We modified the answer selection strategy to make it better suited for opinion summarisation evaluation. We also use different variants of answer verification strategies.

3.1 Answer Selection

In general text summarisation, the datasets used are mostly articles from the news domain, like CNN/DailyMail (Nallapati et al., 2016) and Newsroom (Grusky et al., 2018). Where these articles contain a significant number of named entities(NER) and noun phrases(NP). Different from general text summarisation datasets where the information is contributed heavily by NERs and NPs. For opinion summarisation, there is a limited number of NERs and NPs in reviews. This suggests using NPs alone may not be sufficient to capture opinionated information.

Deutsch and Roth (2020) showed that in addition to the NP, information is expressed in the combination of the verb and its subject child (NSUBJ) and object child (DOBJ). Subrahmanian and Reforgiato (2008) suggested that opinions can also be captured by the combination of adjectives with verbs and adverbs. Our answer selection strategy therefore includes opinion-bearing text spans – adjectives, adverbs, and verbs with their object child and subject child, in addition to NPs. Our experiments evaluating (Section 5) the quality of answers showed that our answer selection strategy can effectively capture the Summarisation Content Units (SCUs) (Nenkova and Passonneau, 2004) in reference summaries. Our further evaluation of the QA pairs shows that the generated QA pairs are of high quality, covering a significant amount of information expressed in SCUs.

3.2 Answer Verification

Previous QA methods (Durmus et al., 2020; Wang et al., 2020; Deutsch et al., 2021) reported the shortcoming of using the F_1 score for answer verification. F_1 score is calculated by using an exact matching of tokens between the answer spans. It works well in an extractive setting but not necessarily in an abstractive scenario. It has the risk of incorrectly penalising a correct answer due to token mismatch.

To overcome the shortcoming of using the F_1 score for answer verification. We propose to leverage LERC (Chen et al., 2020) to verify answers. It

is a learned evaluation metric for reading comprehension to verify the correctness of answers. It was shown to better correlate with human judgements for answer verification by using a more flexible way to evaluate answers. It is achieved by not only comparing the answer spans when generating a score, but also the provided summary and the question.

4 Dataset and Baselines

We conducted experiments to evaluate the QA model and benchmark RunQA against QAEval and other metrics. The dataset we use in our experiments is from Bražinskas et al. (2020b). It was obtained from Amazon product reviews, where 60 products from 4 different categories (15 for each category) were randomly selected. For each product there are 8 source reviews, 1 system generated summary using the Copycat model (Bražinskas et al., 2020b), and 3 reference summaries obtained from Amazon Mechanical Turk [1].

The summary of the dataset is in Table 1. The average number of sentences and words in the reference summaries and reviews is similar. This is because the annotators were instructed to generate summaries of a similar length as the reviews. The candidate summaries have a relatively smaller number of words and sentences generated compared to them.

Document	Avg. No. Words	Avg. No. Sents
Reviews	49.58	3.74
Candidates	34.02	3.13
References	54.54	4.07

Table 1: Statistics of the Amazon review dataset.

Baseline metrics include the lexical overlap-based ROUGE family of metrics (Lin, 2004), embedding-based metrics BERTScore (Zhang et al., 2019) and MoverScore (Zhao et al., 2019), as well as the QA model-based metric QAEval (Deutsch et al., 2021).

5 Evaluation of the QA Model

We conducted experiments to evaluate the effectiveness of our answer selection strategy and the quality of the generated question-answer pairs. To ensure product diversity, we randomly selected two products from each product category. One of the authors manually annotated the SCUs using the reference summaries following the guidelines and using the annotation tool provided by the Pyramid method (Nenkova and Passonneau, 2004) [2].

Token categories in SCUs: The same author applied spaCy[3] to tag tokens in each SCU as categories (e.g. Noun, Verb), and check whether the token is part of a noun phrase(NP). This step aims to examine whether tokens that express information can be successfully captured using NPs only. We put words that do not express information into a separate category and excluded them from our analysis.

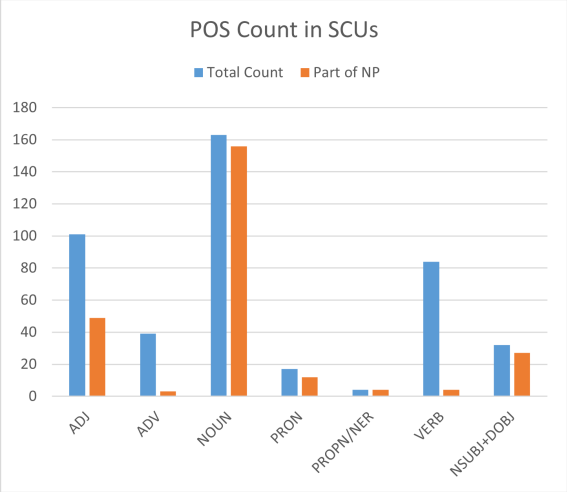

Figure 2: Nouns are a great contributor to NP. There are a significant number of verbs, adverbs, and adjectives presented in the SCUs but not captured by NPs. The number of NSUBJs and DOBJs is not significant due to the structure of SCUs. A limited number of NER suggest that using NP together with NER is not sufficient for review summaries evaluation.

The final result can be found in Figure 2. Not surprisingly, nouns are a great contributor to NPs. Note that there are a significant number of verbs, adverbs, and adjectives presented in the SCUs but not in the NPs. However, since an SCU is similar to a clause but not a full sentence, it is common for an SCU to not contain the subject child (NSUBJ) or the object child (DOBJ) of a verb. The very limited number of proper nouns (NERs) suggests that they do not capture significant information in review summaries. Figure 2 clearly shows that NPs alone cannot capture information in summaries and justifies our proposed approach of selecting answers

[1] https://www.mturk.com/.

[2] http://www1.cs.columbia.edu/~ani/DUC2005/AnnotationGuide.htm.

[3] https://spacy.io/.

based on adjectives, adverbs, and verbs with their subject and objective children as well.

Quality of Generated Question-Answer Pairs: A boxplot of the number of question-answer pairs generated using different answer selection strategies can be found in Figure 3. It is not surprising that our answer selection strategy generates the largest number of QA pairs since we select text spans based on more diverse categories, while using other selection strategies alone generates a limited number of QA pairs. The number of QA pairs generated using NPs only is rather limited, and NERs with NPs do not generate more either, which again suggests that there is a limited number of NERs in review summaries.

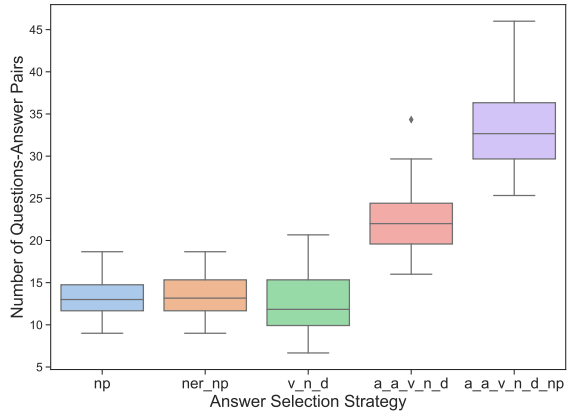

Figure 3: The number of QA pairs generated using different answer selection strategies. Abbreviations: "np" for "noun phrase", "ner_np" for "NER and noun phrase", "v_n_d" for "verb with object and subject child", "a_a_v_n_d" for "adjective, adverb, verb with object and subject child", and "a_a_v_n_d_np" for "adjective, adverb, verb with object and subject child, and noun phrase".

We further evaluated the quality of the generated question-answer pairs. A total of 715 QA pairs were generated for the 8 products in 4 categories. Similar to Deutsch et al. (2021) we calculate the SCU coverage and precision for the QA pairs. QA precision captures the amount of information expressed in the QA pairs that are also in the SCUs. SCU coverage measures the amount of information captured by the SCUs that can also be found in the QA pairs. The results can be found in Table 2. By adopting our answer selection strategy, a significant amount of the information expressed by the SCUs is captured by the generated QA pairs, with 95% coverage and 82% precision. The drop in QA precision is not surprising since we are generating questions more diversely, including information in addition to SCUs.

Strategy	QA Precision	SCUs Coverage
NP	88%	92%
Ours	82%	95%

Table 2: QA precision and SCU coverage by the generated QA pairs for 8 products in 4 categories. Our selection strategy generates QA pairs that have high information overlap with the SCUs.

In summary, we further investigated the quality of the question-answer pairs generated using our proposed strategy. We show that it generates QA pairs that are high quality by covering a significant amount of information expressed in SCUs.

Example questions and answers generated by RunQA can be found in Table 3. In the first example, the first two questions are generated based on NPs. The third question is generated using an adjective. In the first example, the reference answer and the candidate answer do not match for the first two questions, while they match for the third question, which indicates that some information of the reference summary is not captured in the candidate summary.

6 Experiments for Summary Information Quality

Introduced by Koto et al. (2020), coverage and focus are metrics designed to evaluate information in summaries. Coverage(recall) measures the amount of key information expressed in the reference summary that is also captured by the candidate summary. Focus(precision) measures the amount of primary information expressed in the candidate summary that is also true in the reference summary. To compute coverage and focus, we need to gather gold-standard scores by recruiting human annotators. Then calculate the correlation between the human annotations and the scores generated by the metrics.

We followed Koto et al. (2020) and Graham et al. (2017), and used the customised Direct Assessment method to collect human scoring annotations using the Amazon Mechanical Turk. The annotation interface is shown in Figure 4. To control possible bias in annotation, each HIT contained a balanced number of annotations for both coverage and focus using different products (30 different products). On top of the 30 required annotations, each HIT also contained 6 quality control questions. Where 3 are

Reference Summary: This purse is well-designed in terms of appearance, but not in terms of usability and reliability. It is smaller than advertised... **Candidate Summary:** I love this purse! It is a bit smaller than I thought it would be, but I love it! It's a perfect size for me. **Question 1:** This purse is well-designed in terms of what? **Reference Answer**: appearance **Candidate Answer**: perfect size **Question 2:** This purse is well-designed in terms of appearance, but not in terms of what else? **Reference Answer**: usability **Candidate Answer**: NA **Question 3:** How big is it than advertised? **Reference Answer**: smaller **Candidate Answer**: smaller
Reference Summary: This camera bag, constructed as a backpack with padded straps and back is functional and comfortable to wear as well... It is well designed and made of durable materials... **Candidate Summary:** This is the best backpack I have ever owned. It's very comfortable and holds a lot of stuff... **Question 1:** What is this camera bag constructed as with padded straps and back? **Reference Answer**: a backpack **Candidate Answer**: backpack **Question 2:** This camera bag, constructed as a backpack with what type of straps and back is functional and comfortable to wear as well? **Reference Answer**: padded **Candidate Answer**: backpack **Question 3:** What type of material is it well designed and made of? **Reference Answer**: durable **Candidate Answer**: NA

Table 3: Example question and answer pairs generated by RunQA.

exact match summaries that should score 100, and 3 are from random product pair summaries that should score 0.

HITs were restricted to workers from English-speaking countries, with over 10,000 approved HITs and a 98% approval rate. We first collected more than required, then filtered out annotations that failed the quality control tests. It leaves us with an uneven number of annotations per HIT (ranging between 3 and 7). For quality control, we implemented several tests. First, work is only considered if a worker passed 4 of the quality control questions. On top of the distractor questions in each HIT, we further examine workers' time spent on the task and its variation of scores similar to Graham et al. (2017). If the amount of time spent or the variation between the scores is suspiciously low, we disregard all of the worker's annotations. Lastly, the annotations are removed for workers with a Pearson Correlation to other workers(agreement score) of less than 0.2.

After quality control, a mean Pearson Correlation of 0.41 is achieved, with a reasonable average time spent (16.35 minutes) and a quality score (94.74%). Like Koto et al. (2020) and Graham et al. (2017), for annotations pass quality controls we standardise the annotation scores to a z-score of each worker before averaging. This helps remove personal bias introduced by different annotators. Then take an average among workers who completed the same HIT and use the score as the final score for that HIT. We collected both the coverage and focus scores for 180 (60×3) summary pairs.

We use various automatic metrics to generate scores for the candidate summaries based on the reference summaries, and then calculate their correlation with human annotations. Each product has 3 reference summaries, we average both the human and metric scores to one final score for each product. The original BERTScore suggested when comparing against multiple references, the maximum score should be used as the final score. We calculated both the maximum and average BERTScore and found that the average score better correlates with human judgements. Therefore, we use average BERTScore instead of maximum in this paper.

We present the Pearson, Spearman, and Kendall correlations between human annotations for coverage and focus, with various metrics. Results are shown in Tables 4 and 5, all results are significant with p-value <0.01.

The token overlap-based metrics have the weakest correlation with human judgements in both coverage and focus. Context embedding-based met-

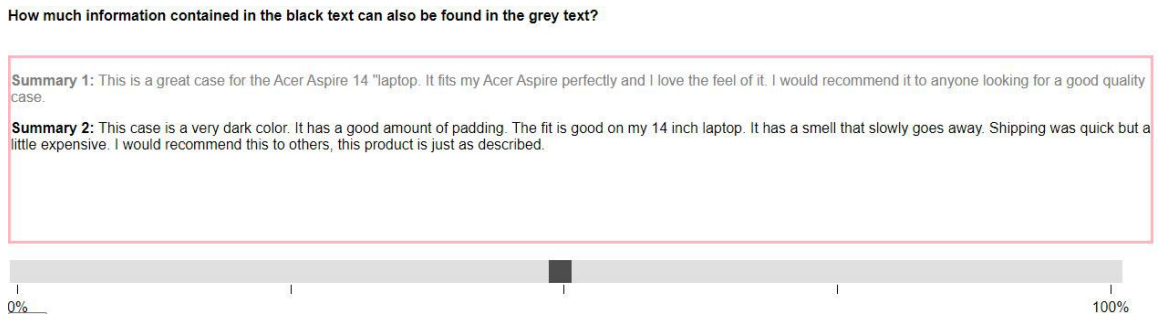

Figure 4: The interface for annotators on the Amazon Mechanical Turk platform.

Metric	r	ρ	τ
ROUGE-1	0.479	0.472	0.310
ROUGE-2	0.413	0.387	0.265
ROUGE-L	0.439	0.403	0.266
MoverScore	0.535	0.471	0.334
BERTScore	**0.599**	**0.549**	**0.398**
QAEval	0.409	0.416	0.290
RunQA (F_1)	0.460	0.484	0.344
RunQA (LERC)	**0.597**	**0.575**	**0.400**

Table 4: Pearson, Spearman and Kendall correlation coefficients of metrics (Coverage)

Metric	r	ρ	τ
ROUGE-1	0.496	0.494	0.339
ROUGE-2	0.525	0.543	0.374
ROUGE-L	0.436	0.388	0.254
MoverScore	0.609	0.597	0.432
BERTScore	**0.651**	**0.645**	**0.470**
QAEval	0.555	0.555	0.409
RunQA (F_1)	0.551	0.654	0.475
RunQA (LERC)	**0.714**	**0.712**	**0.542**

Table 5: Pearson, Spearman and Kendall correlation coefficients for metrics (Focus)

rics show a stronger correlation, especially with BERTScore. Using F_1 to evaluate answers in the QA-based models has a similar performance as the ROUGE family, where we suspect this may be due to the exact match of answers with no consideration of the questions or summaries. Compared with QAEval, correlation improves significantly for RunQA when the answer selection strategy changes to our proposed strategy. RunQA (LERC) has the strongest correlation with human judgements, performs on-par with BERTScore in coverage, and shows the strongest performance in focus.

We further calculated the Pearson correlation for the metrics. The correlation heatmap is shown in

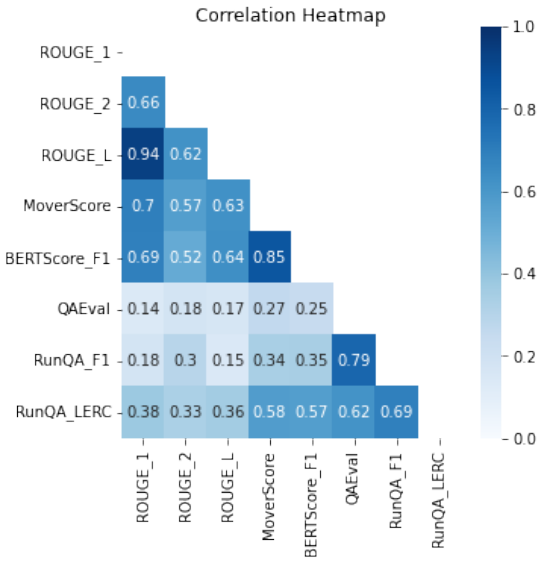

Figure 5: Pearson correlation between metrics.

Figure 5. Similar to Eyal et al. (2019) and Deutsch et al. (2021), we observe that the ROUGE family correlates well with its variants. MoverScore and BERTScore are highly correlated with each other, this is not surprising since BERTScore is a special form of MoverScore. They also have a moderate correlation with the ROUGE variants. Whereas the QA variants have a weak correlation with the ROUGE family. This result suggests that RunQA and QAEval are more likely to evaluate the information expressed in summaries, which is distinct from the lexical overlap in the ROUGE family.

7 Experiments for Ranking Summaries

One of the fundamental requirements for a summarisation metric is to rank summaries and compare the performance of summarisation systems. To simulate the scenario of system summaries with varying quality, we designed a task for ranking summaries generated using the Copycat system (Bražinskas et al., 2020b) and summaries gen-

Metric	Accuracy(%)
ROUGE-1	68.33
ROUGE-2	52.78
ROUGE-L	63.89
BERTScore	54.44
MoverScore	80.56
QAEval	77.78
RunQA (F_1)	82.22
RunQA (LERC)	**91.67**

Table 6: The percentage of each metric that successfully assigns a higher score for the ground-truth summary(human system).

erated by human annotators. We then apply the metrics to calculate the metric scores for human and system summaries. The accuracy that a metric correctly gives human summaries a score higher than that for the Copycat system indicates the reliability of the metric for comparing the human (as an ideal system) and the Copycat system.

Generally, human-generated summaries are of high quality. It is shown in Bražinskas et al. (2020b) that while human ground truth has a relatively close score in fluency compared against the Copycat model, it outperforms all other models significantly in all other dimensions, including Opinion Consensus (measuring whether the summary reflects the common opinions expressed in reviews). Therefore, we would expect the human summary to have better overall quality and should receive a higher score than the system summary.

Previous work (Nenkova and Passonneau, 2004) suggests that individuals tend to both write and pick up information in different ways. For a fair comparison, we compare the summary against different references using each metric. Then pick the higher score because it means the two summaries are closer in terms of information agreement. The same reference is then used in the comparison with the system summary.

We compare the scores generated using different metrics for the human and the Copycat system (Bražinskas et al., 2020b), and count the number of times the human system receives a higher score. Accuracy is calculated by dividing the count by the total number of references (180). The metric with higher accuracy for rating a human summary over a system summary is deemed to be more reliable with a better ability to distinguish between better summaries.

The result is presented in Table 6, BERTScore performs as poorly as the ROUGE family with only 54.44% of correct rankings. RunQA (LERC) performs the best to rank systems, and can distinguish the better quality human summary from the system-generated summary over 90% of the time. We suspect this is because RunQA ranks summaries based on answerable questions hence evaluating the information quality. The embedding-based metrics examine distance between tokens. If the layout of the summaries is similar but express opposite opinions, scores will be similar, which makes it hard to rank summaries based on information quality. As discussed in (Tay et al., 2019), ROUGE is not sensitive to opinion mismatch, which explains its poor performance.

8 Limitations

While our study has shown that RunQA is a better metric for opinion summarisation evaluation. There are some limitations with our research. First, we did the experiments using only one summarisation system. It would be more assuring if we explored other summarisation models.

The dataset (Bražinskas et al., 2020b) in our experiments is the only publicly available dataset with a significant number of products and multiple ground-truth summaries. The dataset is from the product review domain and is not representative of the other domains, such as restaurant or movie reviews. Using multiple datasets and summarisation models would give a better picture in terms of human correlation in different domains.

9 Conclusion

We proposed RunQA, which uses the Question-Answering model to evaluate review summarisation. We proposed to identify answers based on opinion-bearing token categories to generate QA pairs. Experiments on a public Amazon review summary dataset show that RunQA correlates well with human judgements for evaluating the amount of salient opinion captured in the candidate summary against the reference summary. RunQA is also more reliable than existing metrics in the literature for ranking summaries.

RunQA has shown high potential when used for opinion summarisation evaluation for opinion quality. Our future work will explore applying RunQA for review summarisation evaluation in other domains.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020a. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020b. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6521–6532, Online. Association for Computational Linguistics.

Eric Chu and Peter Liu. 2019. Meansum: a neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*, pages 1223–1232. PMLR.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Daniel Deutsch and Dan Roth. 2020. Understanding the extent to which summarization evaluation metrics measure the information quality of summaries. *arXiv preprint arXiv:2010.12495*.

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.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. 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 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. Supert: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

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)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2020. Ffci: A framework for interpretable automatic evaluation of summarization. *arXiv preprint arXiv:2011.13662*.

Lun-Wei Ku, Yu-Ting Liang, Hsin-Hsi Chen, et al. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 100107, pages 1–167.

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*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Sujian Li, and Houfeng Wang. 2012. Entity-centric topic-oriented opinion summarization in twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 379–387.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Ça glar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *CoNLL 2016*, page 280.

Ani Nenkova and Rebecca J Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, pages 145–152.

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*, pages 1925–1930.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. 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 of the Association for Computational Linguistics*, pages 311–318.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Venkatramana S Subrahmanian and Diego Reforgiato. 2008. Ava: Adjective-verb-adverb combinations for sentiment analysis. *IEEE Intelligent Systems*, 23(4):43–50.

Wenyi Tay. 2019. Not all reviews are equal: Towards addressing reviewer biases for opinion summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 34–42, Florence, Italy. Association for Computational Linguistics.

Wenyi Tay, Aditya Joshi, Xiuzhen Jenny Zhang, Sarvnaz Karimi, and Stephen Wan. 2019. Red-faced rouge: Examining the suitability of rouge for opinion summary evaluation. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 52–60.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020.

Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020. Unsupervised reference-free summary quality evaluation via contrastive learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3612–3621.

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. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.