Red-faced ROUGE: Examining the Suitability of ROUGE for Opinion Summary Evaluation

Wenyi Tay[1,2], **Aditya Joshi**[2], **Xiuzhen Zhang**[1], **Sarvnaz Karimi**[2] **and Stephen Wan**[2]
[1]RMIT University, Australia
[2]CSIRO Data61, Australia
{wenyi.tay, xiuzhen.zhang}@rmit.edu.au
{Aditya.Joshi, Sarvnaz.Karimi, Stephen.Wan}@data61.csiro.au

Abstract

One of the most common metrics to automatically evaluate opinion summaries is ROUGE, a metric developed for text summarisation. ROUGE counts the overlap of word or word units between candidate summaries and reference summaries. This formulation treats all words in the reference summary equally. In opinion summaries, however, not all words in the reference are equally important. Opinion summarisation requires to correctly pair two types of semantic information: (1) opinion target, or aspect; and, (2) polarity of candidate and reference summaries. We investigate the suitability of ROUGE for evaluating opinion summaries of online reviews. We design three experiments to evaluate the behaviour of ROUGE for opinion summarisation on the ability to capture aspect and polarity. We show that ROUGE cannot distinguish opinion summaries of the same or opposite polarities for the same aspect. Moreover, ROUGE scores have significant variance under different configuration settings. As a result, we present three recommendations for future work on evaluating opinion summaries.

1 Introduction

Popular e-commerce websites allow users to express their opinion about products or services in the form of reviews. An opinion is formally defined as a combination of *aspect* (an attribute of the product or service as the opinion target, expressed through aspect words), and *sentiment polarity* (either positive or negative, expressed through opinion words) (Liu, 2012). The opinion expressed in online reviews potentially helps prospective buyers to make decisions. Given the large volume of reviews, it is time-consuming and often impractical for a user of these websites to read all reviews pertaining to the set of products that they are considering to purchase. This makes opinion summarisation important because it allows users to obtain aggregate key opinions about the product or service based on its reviews. Given the value of opinion summaries, automatic opinion summarisation is an active area of research with the focus of producing high-quality opinion summaries.

The quality of an opinion summary would ideally be evaluated by human annotators. For example, annotators may read a summary and rate it according to quality measures such as informativeness, ability to capture sentiment polarity, coherence and redundancy (Angelidis and Lapata, 2018). However, human evaluation is resource-intensive and not scalable. This motivates automatic evaluation. In the case of automatic evaluation, for a set of product reviews, reference summaries are written by human experts apriori. *Reference summaries* are the ground truth summaries against which *candidate summaries* to be evaluated.

Using reference and candidate summaries, automatic evaluation of opinion summaries adopts metrics from text summarisation (Lin, 2004) and machine translation (Papineni et al., 2002; Lavie and Denkowski, 2009). We focus on the most frequently reported metric for opinion summarisation, Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) and leave the analysis of other evaluation metrics to future studies. ROUGE counts the overlap of word or word units between the candidate summary and reference summary with respect to the word units in the reference summary. A higher ROUGE score means a larger overlap of word units while a lower ROUGE score means a smaller overlap of word units. The ROUGE scores can then be used as a criterion to compare summaries and systems.

An opinion is a combination of aspect and sentiment, thus, the evaluation must assess both. ROUGE treats the contribution of all matched word units in the reference summary to the ROUGE score equally. This may not hold in the

case of opinion summaries. For example, a lack of match to aspect terms in the reference summary may be due to the different ways people refer to the same aspect, or that it does not contain that aspect and thus the opinion is not present. The lack of a match to sentiment-bearing terms can mean there is no opinion present in the summary, the opinion is consistent but expressed differently or the opinion is opposite to the reference summary. That ROUGE does not differentiate the reasons for the lack of a word match or the reasons for mismatch has implications to evaluate opinions summaries. To date, opinion summarisation has been considered a special case of text summarisation. Therefore, the popularity of ROUGE for opinion summarisation seems intuitive. However, the distinction between opinion summarisation and text summarisation warrants a critical examination of the utility of ROUGE for opinion summarisation. Our research question is:

'*Can ROUGE scores be used to correctly compare summaries to ensure that the candidate summary is accurate to the opinion aspect and polarity in the reference summary?*'

This paper makes two-fold contributions: (1) Through experiments, we demonstrate that ROUGE is not able to accurately evaluate the opinions in the candidate summary against the reference summary[1]; and (2) Our discussion provides three recommendations for further research on opinion summary evaluation.

2 Related Work

Early work in opinion summarisation conducted their evaluation using metrics other than ROUGE. Pang and Lee (2004), an early work in extractive opinion summarisation, use sentence-level accuracy. Lerman et al. (2009) pre-date the existence of opinion summarisation datasets. Therefore, they use human evaluation for their systems. Pitler et al. (2010) propose an automatic metric for summarisation. This metric captures linguistic quality using a set of features. They conclude that syntactic features are the best indicators for linguistic quality of summaries. Following the availability of datasets with opinion summaries, ROUGE could be used for opinion summarisation. However, its value has been understood to be limited for the evaluation of opinion summarisation. Jayanth et al. (2015) observe that ROUGE is influenced by topic terms more than sentiment terms. Therefore, they report two metrics: ROUGE scores and sentiment correlation. Mackie et al. (2014) show that, for microblog summarisation, ROUGE does not correlate with human judgment as well as a more naïve indicator: fraction of topic words. In addition, limitations of ROUGE to evaluate text summarisation have also been reported. Conroy and Schlesinger (2008) show that ROUGE may not correlate well with human evaluation for text summarisation, and needs to be combined with human scores. More recently, Graham (2015) present an extensive comparison of 192 variants of ROUGE, and show that the metrics have contrasting conclusions. Table 1 summarises key studies and the choice of automatic evaluation metrics used for opinion summarisation.

Despite the limitations, ROUGE is the most popular metric for opinion summarisation (Moussa et al., 2018). It continues to be used as an automatic evaluation in recent papers either on its own (Anchiêta et al., 2017) or in combination with other automatic metrics such as METEOR (Amplayo and Lapata, 2019). Angelidis and Lapata (2018) report ROUGE for multi-document opinion summarisation.

Alternatives to ROUGE have been proposed. Kabadjov et al. (2009) use sentiment intensity to measure sentiment summarisation. Kunneman et al. (2018) use gold standard summaries available in the forum as reference summaries, and report precision, recall and F_1 scores. Poddar et al. (2017) use a combination of lexical and sentiment similarity to capture sentiment-aware similarity between sentences.

We note that past work states the limitations of ROUGE as a part of the discussion of the results of their proposed systems, while examining these limitations is the focus of our work.

3 ROUGE

ROUGE measures content coverage of candidate summaries against reference summaries (Lin, 2004). Different variants of ROUGE have been proposed. For example, ROUGE-N, ROUGE-L and ROUGE-S count the number of overlapping units of n-gram, word sequences, and word pairs between the candidate summary and the reference

[1]Although the current analysis focuses on ROUGE for evaluating opinion summaries, the limitations of using word matching for evaluation is also a problem faced by text summarisation and text generation.

	Dataset	Task	ROUGE	Others
Ganesan et al. (2010)	Opinosis	Abstractive	R-1,R-2,R-SU4	No
Jayanth et al. (2015)	Movie	Abstractive	R-1,R-2	Senti Corr
Wang and Ling (2016)	RottenTomatoes	Abstractive	R-SU4	BLEU, METEOR
Angelidis and Lapata (2018)	Oposum	Extractive	R-1,R-2,R-L	No
Kunneman et al. (2018)	ProductReviews	Abstractive	No	Precision, Recall and F_1
Amplayo and Lapata (2019)	RottenTomatoes	Abstractive	R-1 R-2,R-L,R-SU4	METEOR

Table 1: Summary of automatic evaluation metrics used to evaluate opinion summaries.

Reference: The rooms were neat and clean.			**Summary1:** Clean room.		**Summary2:** The rooms were dirty.			
Configuration	**R-1**	**R-2**	**R-L**	**R-SU4**	**R-1**	**R-2**	**R-L**	**R-SU4**
None	0.250	0.000	0.250	0.087	0.600	0.500	0.600	0.400
Stemming	0.500	0.000	0.250	0.174	0.600	0.500	0.600	0.400
StopWordRemoval	0.400	0.000	0.400	0.222	0.400	0.000	0.400	0.222
StopWordRemoval+Stemming	0.800	0.000	0.400	0.444	0.400	0.000	0.400	0.222

Table 2: F_1 score of ROUGE metrics under different configurations.

summaries respectively. The formula for ROUGE-N is shown in equation 1, where $gram_n$ is the choice of n-gram and S is the reference summary:

$$\text{ROUGE-N} = \frac{\sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{gram_n \in S} Count(gram_n)}. \quad (1)$$

The following considerations are important when using ROUGE:

1. **Multiple reference summaries**: A candidate summary with highest score for each pairwise evaluation of the candidate summary against each reference summary will be the ROUGE score for the summary. In this paper, we assume one reference summary.
2. **Pre-processing configurations**: Different pre-processing configurations can be taken into account. These typically include stemming and stop word removal. There are no recommended or commonly agreed pre-processing configurations. In this paper, we compare multiple combinations of these pre-processing configurations.
3. **Other configurations**: Although ROUGE is a recall-based metric, there is an option to report precision and F_1 scores with each ROUGE metric. The precision score takes the overlapping word units with reference to the word units of the Candidate summary. The F_1 score is the harmonic mean of the recall and precision score. In our work, we investigate the different configurations of ROUGE for opinion summary evaluation.
4. **Comparing different systems**: ROUGE scores can be used to compare summarisation systems by taking the mean or median of all the summaries generated by the system. Should the ROUGE score at summary level be incorrect, the error propagates to the system level. We focus on the ROUGE scores at the summary level.

To demonstrate how ROUGE is computed, we consider the following hypothetical examples:
Reference summary: The rooms were neat and clean.
Candidate summary 1: Clean room.
Candidate summary 2: The rooms were dirty.

The opinion in Candidate summary 1 is consistent with the Reference summary since the two refer to clean rooms. Candidate summary 2 gives an opinion that is opposite to the Reference summary because it states that the rooms were dirty. Intuitively, Candidate summary 1 should be evaluated better as compared to Candidate summary 2.

We report the ROUGE scores of ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L) and ROUGE-SU4 (R-SU4) for these examples. Table 2 shows the ROUGE scores for various ROUGE metrics with different configurations for our example summaries. We see that for configurations "None" and "Stemming", all ROUGE metrics for Candidate summary 2 are higher than Candidate summary 1. In the case of "Stop-

No. of gold standard summaries	223
Min No. of Words in a summary	3
Max No. of Words in a summary	62
Average No. of Words in a summary	16.7
Median No. of Words in a summary	15

Table 3: Statistics of Opinosis Dataset.

WordRemoval", both summaries are the same. However, for "StopWordRemoval+Stemming", ROUGE metrics score Candidate summary 1 higher than Candidate summary 2. This demonstrates that the selection of the pre-processing configuration for ROUGE metrics may affect the scoring of summaries thus affecting the comparison of summaries.

The previous motivating example highlights that opinion summary evaluation requires a different notion of content coverage compared to text summarisation. Content coverage for opinion summaries is not just matching words. Opinion summary evaluation requires differential comparison of two groups of words, the aspect terms and sentiment-bearing words. We design three experiments that create summaries to reflect semantic and sentiment variability. They are described in the forthcoming section.

4 Experiment and Results

Opinosis is a opinion summary dataset by Ganesan et al. (2010). It contains 51 documents, where each document is a collection of opinions from online reviews on one aspect of hotels, cars and products. Examples of aspects are service for hotels, mileage of cars and size of netbooks (Note that aspects are called topics in the Opinosis dataset (Ganesan et al., 2010).) Each document is associated with three to five gold standard summaries. The gold standard summaries are created by human annotators by asking them to summarise the major opinions in the document. We observe duplicates in the gold standard summaries. After removing duplicates, there are 223 gold standard summaries left. Some key statistics of the dataset are listed in Table 3.

4.1 Summary Triplet Experiment

Our first experiment investigates a trivial case: "How does ROUGE respond when evaluating candidate summaries of similar or different aspects to the reference summary?"

We investigate this problem using summary triplets. Each triplet is made up of: (1) a reference summary (*Reference*); (2) a candidate summary of the same aspect (*Summ-SameAsp*); and (3) a candidate summary of a different aspect (*Summ-DiffAsp*). We begin by taking one gold standard summary as *Reference*. *Summ-SameAsp* is a randomly selected gold standard summary with the same aspect as *Reference*. For *Summ-DiffAsp*, we randomly select a summary with a different aspect from *Reference*. We repeat this process for every gold standard summary in the dataset. We have a total of 223 summary triplets.

Since the same aspect can be referred to by different words, we wish to avoid the bias from the same (different) aspect terms in the second (third) summary for the ROUGE metrics. Therefore, we mask the aspect terms in summaries[2]. Table 4 shows two examples of summary triplets. Observe that the terms in the *Summ-DiffAsp* are generally different from those in *Summ-SameAsp*.

Table 5 shows the proportion of triplets that *Summ-SameAsp* is scored higher than *Summ-DiffAsp*; higher values indicate better performance of ROUGE for ranking candidate summaries. The Recall score of ROUGE variants perform reasonably well with around 60% "accuracy" except for R-2. This result confirms our observation that there are few overlapping words between the candidate summary and reference summary when they are of different aspects.

We closely analyse the poor performance of R-2 and we found that it is due to the many ties in the scores of *Summ-SameAsp* and *Summ-DiffAsp*. In particular, many candidate summaries have a R-2 score of zero, as shown in Figure 1. When ROUGE scores are zero for both candidate summaries, scores are no longer meaningful for evaluation of candidate summaries.

We also observe that, with "Stemming", the proportion is higher than otherwise. One possible explanation is that "Stemming" relaxes the exact word match requirement to allow matching of different word forms.

We further examine how the choice of Recall, Precision or F_1 score affects the suitability of using ROUGE to compare summaries. In Table 5, we report the proportion of triplets that score

[2]For two aspects, summaries do not contain the given aspect terms. For the aspect "accuracy", we mask the word "accurate" and for aspect "eyesight-issues", we mask the word "eyes".

Reference	Summ-SameAsp	Summ-DiffAsp	R-1$_{\text{Summ-SameAsp}}$	R-1$_{\text{Summ-DiffAsp}}$
Great ⟨performance⟩ and handling. ⟨Performance⟩, styling and quality have good value for money.	Adequate ⟨performance⟩, nice looks, long distance cruiser. Overall ⟨performance⟩ good, poor engine ⟨performance⟩, gas mileage 22 highway and poor comfort level.	⟨Price⟩ was good. Best ⟨prices⟩ on other websites than Holiday Inn.	0.069	0.125
The ⟨rooms⟩ were very neat and clean.	The ⟨rooms⟩ are clean, large and comfortable. Not the most modern decor, however.	The Best Western in San Francisco is a decent, inexpensive option for one's stay in the city. It's ⟨location⟩ is terrific , being near many attractions, and the rooms, while small, are clean.	0.222	0.0.95

Table 4: Two examples of summary triplets. *Summ-SameAsp* is a randomly selected gold standard summary on the same aspect. whereas *Summ-DiffAsp* is on a different aspect. Words in angle bracket are the aspect terms we masked. We report the R-1 F_1 score with stemming and stop word removal.

	Recall				F_1 score			
Configuration	**R-1**	**R-2**	**R-L**	**R-SU4**	**R-1**	**R-2**	**R-L**	**R-SU4**
None	0.659	0.323	0.578	0.677	0.758	0.345	0.744	0.767
Stemming	0.682	0.341	0.596	0.695	0.771	0.359	0.744	0.776
StopWordRemoval	0.610	0.108	0.592	0.610	0.646	0.108	0.628	0.646
StopWordRemoval+Stemming	0.641	0.139	0.632	0.641	0.677	0.139	0.677	0.677

Table 5: Proportion of 223 summary triplets that *Summ-SameAsp* is scored higher than *Summ-DiffAsp*, by Recall and F_1 score, when the aspect terms are masked.

Summ-SameAsp higher than *Summ-DiffAsp*, using Recall and F_1 score. When using Recall score, the proportion of correctly assessed triplets is lower. Using F_1 score, the proportion is higher suggests that the Precision score plays a part in the comparison of candidate summaries of two different aspects and in a way controls for the different lengths in the candidate summaries.

There are three learning points to this experiment: (1) ROUGE gives a low score to candidate summary of a different aspect to the reference summary; (2) ROUGE-N score decreases as n-gram increases. It is possible that ROUGE-N scores are zero. Hence, is useful to plot the distribution of ROUGE scores; and, (3) Results suggest that "Stemming" increases ROUGE's ability to compare summaries.

4.2 Same Polarity Triplet Experiment

We had two annotators read all 223 gold standard summaries for 51 aspects and assign either a positive or negative sentiment polarity to each summary. When there is a conflict between the two assigned labels, a third annotator decides if the summary is positive or negative. Out of the 51 aspects, for 38 aspects (74.5%) the gold summaries of each aspect were consistent in their polarity whereas for 13 aspects (24.6%) the gold summaries were opposite.

Our second experiment is on the 38 aspects where all gold summaries have the same polarity. We design the experiment in a controlled way to study how ROUGE ranks candidate summaries containing same and opposite polarities compared to the same reference summary. Using the idea of a triplet summary as before, we create a triplet consisting of: (1) a reference summary (*Reference*); (2) a candidate summary that is consistent in aspect and sentiment polarity (*Summ-Syn*); and, (3) a candidate summary of the same aspect but opposite sentiment polarity (*Summ-Ant*). From the summary triplets we generated in the previous section, we use *Reference* and *Summ-SameAsp* summaries. By replacing the sentiment-bearing words of *Summ-SameAsp* with its synonym or antonym, we generate two versions of the same summary. We have a candidate summary that is consistent in aspect and sentiment polarity, *Summ-Syn*, and an-

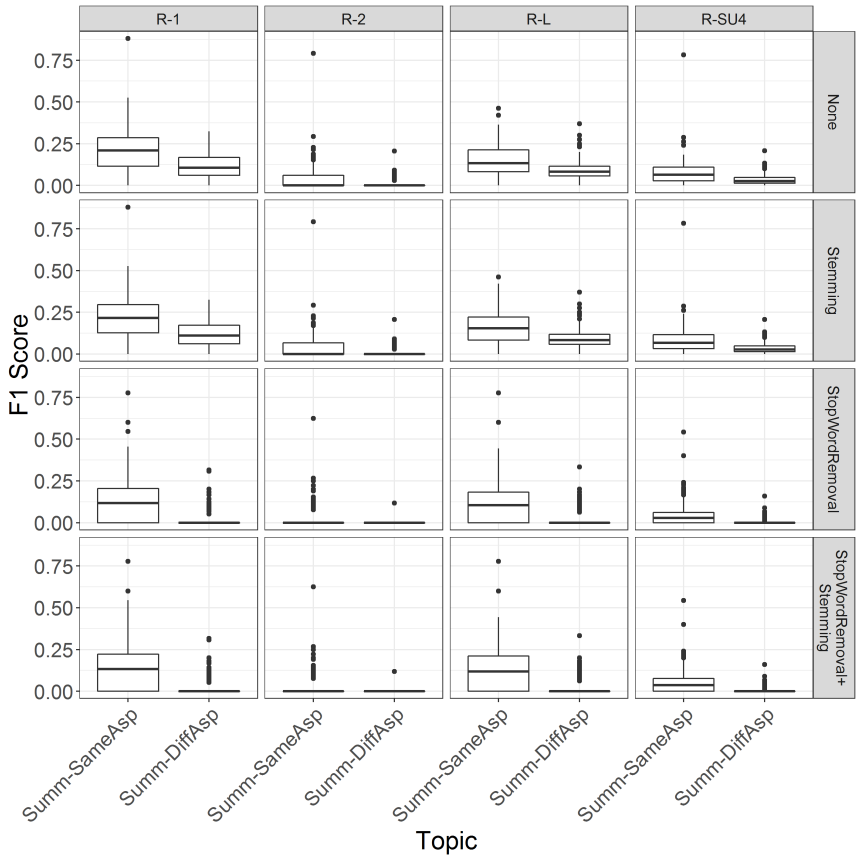

Figure 1: Boxplot of ROUGE scores show that scores for *Summ-DiffAsp* are generally lower than scores for *Summ-SameAsp*. Also, R-2 scores with "StopWordRemoval" are close to zero for candidate summaries which makes it less meaningful to be used to compare summaries.

other summary that is of same aspect but opposite polarity, *Summ-Ant*. This forms the second and third summaries of the triplet. This experiment controls for all the matching of the other words except for the sentiment-bearing words. Hence, we can study the impact of matching sentiment-bearing words in the reference summary.

To generate the synonym and antonym version of a summary, we first identify the sentiment-bearing words in the summary. A sentiment-bearing word is an adjective, adverb or verb and its lemmatised word form contains a sentiment score in SentiWordNet (Baccianella et al., 2010). The pre-processing steps of part-of-speech tagging, lemmatisation and looking it up in SentiWordnet was performed through python's NLTK package (Bird et al., 2009). We obtain synonyms and antonyms[3] from Wiktionary using the python package wiktionaryparser. Table 6 reports the proportion of sentiment-bearing words present in the gold standard summaries and Table 7 shows examples of the synonyms and antonyms from Wiktionary.

Intuitively, *Summ-Syn* is accurate to *Reference* summary. As such, we expect *Summ-Syn* to be evaluated as a better summary over *Summ-Ant*. But, based on the ROUGE formula, we expect similar ROUGE scores for both candidate summaries.

Not all triplets have a synonym and antonym summary due to the nature of the synonym and antonym extraction method. We exclude triplets where there are no synonym and antonym summaries. We are left with 104 summary triplets with synonym and antonym summaries with at least one sentiment-bearing word replaced. On average, 0.124 of the summary is replaced by antonyms or synonyms. Table 8 shows two examples of summary triplet and Table 9 shows the proportion of triplets that ROUGE scored both summaries the same score.

From Table 9, most triplets have the same

[3]We also experimented with WordNet (Fellbaum, 1998) to get synonym and antonym of a sentiment-bearing word, however, the pairs we obtained from WordNet had lower coverage than Wiktionary.

	Proportion	Examples
Adjectives	0.122	easy, clean, friendly
Adverbs	0.050	not, very, too
Verbs	0.110	is, like, was

Table 6: Proportion of sentiment-bearing words of all words in gold standard summaries according to their part-of-speech tag.

Word	Synonym	Antonym
small	little	large
large	big	small
exceptional	excellent	ordinary
inferior	bad	superior
worse	unfavorable	good

Table 7: Five examples of sentiment-bearing words with its synonym and antonym.

score as expected from our understanding of the ROUGE formula. ROUGE scores cannot be used to differentiate summaries that are accurate to the reference summary in terms of sentiment polarity.

4.3 Opposite Polarity Triplet Experiment

Our third experiment is on the 13 aspects where gold summaries are not consistent in sentiment polarity. For example, for the topic "buttons_amazon_kindle", there are 1 negative summary and 3 positive summaries. We create a triplet consisting of: (1) a reference summary (*Reference*); (2) a candidate summary that is consistent in aspect and sentiment polarity (*Summ-SamePol*); and, (3) a candidate summary of the same aspect but opposite sentiment polarity (*Summ-DiffPol*). We took all possible combinations with the annotated gold standard summaries. We have a total of 142 summary triplets. An example of the triplet is shown in Table 10.

We report the proportion of the 142 summary triplets where *Summ-SamePol* is scored higher than *Summ-DiffPol* in Table 11. R-2 is excluded as the R-2 scores of both candidate summaries are mostly zero, which are not meaningful to compare summaries. We observe that in general, the proportion of triplets where ROUGE scores the second summary higher is lower than 50%. This suggests that ROUGE is not able to correctly rank candidate summaries of the same polarity with the reference summary. Also, we observe that configurations with "StopWordRemoval" is always lower than the configurations without. From all experiments, the inclusion of "StopWordRemoval" often reduces the effectiveness of the the ability to use ROUGE scores to compare candidate summaries.

5 Discussion

Our empirical analysis for examining whether ROUGE is suitable for evaluating opinion summaries leads us to three suggestions for future studies for automatic evaluation in opinion summarisation:

1. The configurations for ROUGE can change or reverse the order of scores of summary. We observe that F_1 scores appear to compare summaries better than Recall. Also, "StopWordRemoval" seems to reduce the ability of ROUGE scores for comparing summaries for our dataset. Including "Stemming" often improve the ability to compare candidate summaries for our dataset. Hence, when reporting ROUGE scores, in addition to reporting ROUGE variants, we recommend reporting the configurations under which ROUGE was computed.
2. ROUGE scores will be low when candidate summary is of a different aspect from the reference summary. This is because opinions for different aspects are described by different sets of words. As such, there is little word overlap which leads to low ROUGE scores. Hence, for improvements to the opinion summary evaluation, we recommend checking for a match of the aspect in candidate and reference summary as a differentiating criteria.
3. It is not possible to infer from ROUGE scores if the candidate summary is accurate to the reference especially for sentiment polarity. ROUGE requires an exact match of the sentiment-bearing words in the reference summary. But reviewers express opinions differently which can result in the lack of match of sentiment-bearing words. We recommend sentiment agreement of candidate and reference summaries as another criteria for evaluation.

6 Conclusions

ROUGE is a popular metric for automatic evaluation of opinion summarisation. However, using ROUGE as a means to measure content coverage

Reference	Summ-Syn	Summ-Ant	R-1$_{\text{Summ-Syn}}$	R-1$_{\text{Summ-Ant}}$
Great ⟨performance⟩ and handling. ⟨Performance⟩, styling and quality have good value for money.	Adequate ⟨performance⟩, charming looks, long distance cruiser. Overall ⟨performance⟩ good, impoverished engine ⟨performance⟩, gas mileage 22 highway and impoverished comfort level.	Adequate ⟨performance⟩, horrible looks, long distance cruiser. Overall ⟨performance⟩ good, rich engine ⟨performance⟩, gas mileage 22 highway and rich comfort level.	0.069	0.069
The ⟨rooms⟩ were very neat and clean.	The ⟨rooms⟩ are clean, big and comforting. Not the most contemporary decor, however.	The ⟨rooms⟩ are clean, small and comfortless. Not the most ancient decor, however.	0.222	0.222

Table 8: Two examples of summary triplet. *Summ-Syn* is a synonym version of a gold standard summary. *Summary-Ant* is an antonym version of a gold standard summary. We report the R-1 F_1 score with stemming and stop word removal. The words that are replaced in the original summary are underlined.

Configuration	R-1	R-2	R-L	R-SU4
None	0.952	1.000	0.952	0.952
Stemming	0.933	1.000	0.942	0.933
StopWordRemoval	0.952	1.000	0.942	0.952
StopWordRemoval+Stemming	0.913	0.990	0.904	0.913

Table 9: Proportion of 104 summary triplets with same ROUGE scores for *Summ-Syn* and *Summ-Ant*.

Reference	Summ-SamePol	Summ-DiffPol	R-1$_{\text{Summ-SamePol}}$	R-1$_{\text{Summ-DiffPol}}$
New ⟨buttons⟩ are easy to use and effective. No more accidental ⟨button⟩ presses. ⟨Buttons⟩ make navigation easy.	Magical five way ⟨button⟩. Next page ⟨button⟩ on both side of kindle. No reset ⟨button⟩.	It is not user friendly and the ⟨buttons⟩ are not easily pressed.	0.000	0.118

Table 10: An examples of summary triplet. *Summ-SamePol* and *Summ-DiffPol* are gold standard summaries of the same aspect as *Reference* with same and different polarity respectively. We report the R-1 F_1 score with stemming and stop word removal.

Configuration	R-1	R-L	R-SU4
None	0.479	0.465	0.493
Stemming	0.479	0.465	0.500
StopWordRemoval	0.430	0.387	0.437
StopWordRemoval+Stemming	0.451	0.394	0.444

Table 11: Proportion of 142 summary triplets where *Summ-SamePol* is scored higher than *Summ-DiffPol*.

is not sufficient for the evaluation of opinion summaries. The word count overlap is not an indicator of accurate opinion summarisation. Our experiments simulate scenarios where inaccurate summaries are automatically generated. We observe that ROUGE is unable to differentiate summaries that are accurate and summaries that are inaccurate. For future work, we will investigate opinion summaries that contain multiple opinions. Based on the learning points from the investigation, we aim to propose a new metric that incorporates semantic similarity in terms of opinion target and opinion polarity.

Acknowledgments

We thank the anonymous reviewers for their thorough and insightful comments. Wenyi is supported by an Australian Government Research Training Program Scholarship and a CSIRO Data61 Top-up Scholarship.

References

Reinald Kim Amplayo and Mirella Lapata. 2019. Informative and controllable opinion summarization.

arXiv preprint arXiv:1909.02322.

Rafael Anchiêta, Rogerio Figueredo Sousa, Raimundo Moura, and Thiago Pardo. 2017. Improving opinion summarization by assessing sentence importance in on-line reviews. In *Proceedings of the Brazilian Symposium in Information and Human Language Technology*, pages 32–36, Uberlândia, Brazil.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the conference on International Language Resources and Evaluation*, Valletta, Malta.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

John M Conroy and Judith D Schlesinger. 2008. CLASSY and TAC 2008 Metrics. In *Proceedings of the Text Analysis Conference*, Gaithersburg, MD.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions. In *Proceedings of the International Conference on Computational Linguistics*, pages 340–348, Beijing, China.

Yvette Graham. 2015. Re-evaluating Automatic Summarization with BLEU and 192 Shades of ROUGE. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal.

Jayanth Jayanth, Jayaprakash Sundararaj, and Pushpak Bhattacharyya. 2015. Monotone submodularity in opinion summaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 169–178, Lisbon, Portugal.

Mijail Kabadjov, Alexandra Balahur, and Ester Boldrini. 2009. Sentiment intensity: Is it a good summary indicator? In *Language and Technology Conference*, pages 203–212. Springer.

Florian Kunneman, Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2018. Aspect-based summarization of pros and cons in unstructured product reviews. In *Proceedings of the International Conference on Computational Linguistics*, pages 2219–2229, Santa Fe, NM.

Alon Lavie and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105–115.

Kevin Lerman, Sasha Blair-Goldensohn, and Ryan McDonald. 2009. Sentiment summarization: evaluating and learning user preferences. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 514–522, Athens, Greece.

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

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Stuart Mackie, Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2014. On choosing an effective automatic evaluation metric for microblog summarisation. In *Proceedings of the Information Interaction in Context Symposium*, pages 115–124.

Mohammed Elsaid Moussa, Ensaf Hussein Mohamed, and Mohamed Hassan Haggag. 2018. A survey on opinion summarization techniques for social media. *Future Computing and Informatics Journal*, 3(1):82–109.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, page 271, Barcelona, Spain.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 544–554, Uppsala, Sweden.

Lahari Poddar, Wynne Hsu, and Mong Li Lee. 2017. Author-aware Aspect Topic Sentiment Model to Retrieve Supporting Opinions from Reviews. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 472–481, Copenhagen, Denmark.

Lu Wang and Wang Ling. 2016. Neural Network-Based Abstract Generation for Opinions and Arguments. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, CA.