Evaluation of Review Summaries via Question-Answering

Nannan Huang and Xiuzhen Jenny Zhang | RMIT University
Contact: s3754491@student.rmit.edu.au, xiuzhen.zhang@rmit.edu.au

Background

- Opinion summarisation is the task of compressing multiple opinionated documents into a single concise summary reflecting key information expressed
- Advancement in model development:
  - ❖ From: Extractive (copy and paste key phrases)
  - ❖ To: Abstractive (paraphrasing)
- Evaluation metrics lag behind
  - ❖ ROUGE[1] still the only automatic metric being used in recent studies
  - ❖ Problems:
    - ➢ Not evaluating opinion consensus[2]
    - ➢ Not suitable for opinion summarisation evaluation[3]
    - ➢ Not suitable for abstractive summarisation evaluation[4]
- Review summaries should be evaluated based on opinions
- Existing metrics are not evaluating information[2]
- QA-based metrics are proven to evaluate information[5]

Goal: Develop a metric that evaluates opinion summarisation systems based on opinion consensus

Objectives: Improve the QA-based metric to more effectively evaluate the opinions expressed in the review summaries

Methodology

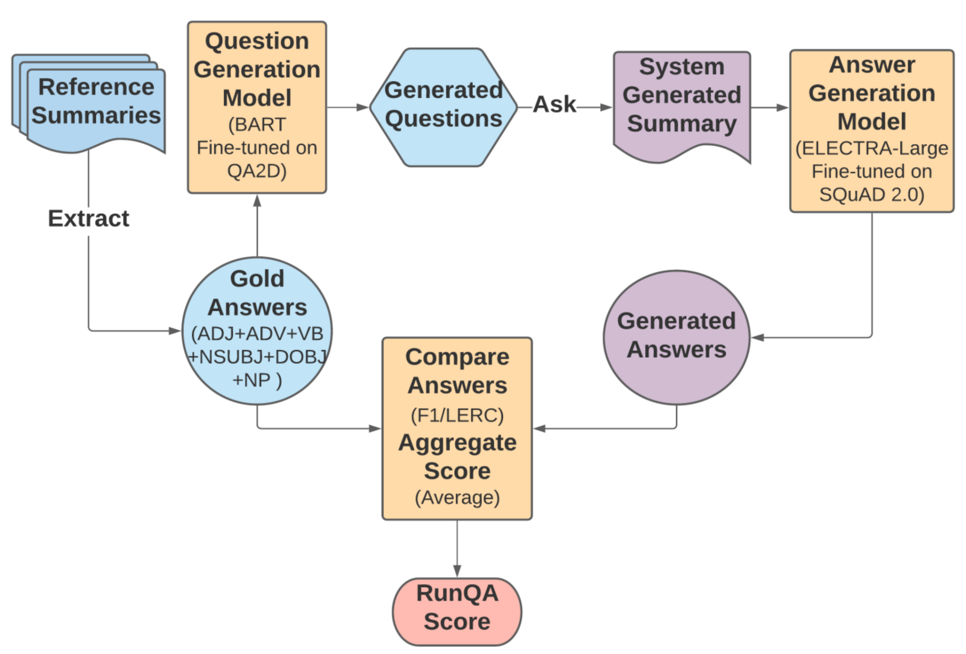

Figure 1: RunQA: Review Summaries Evaluation via Question-Answering model architecture.

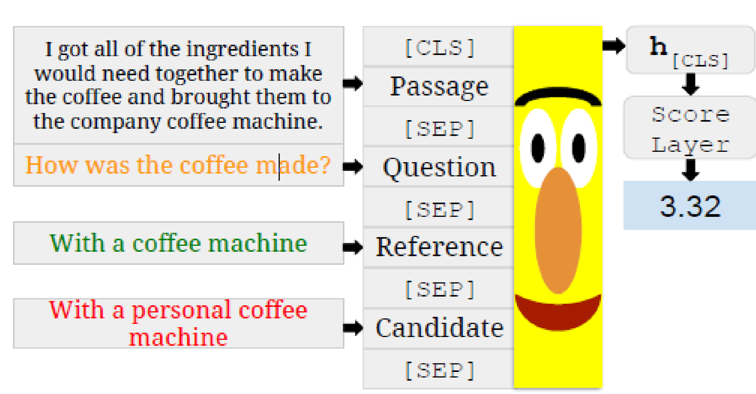

Figure 2: Architecture of LERC[6]

- Not comparing text at a surface level
  - ❖ Extract "ground-truth"
  - ❖ Generate questions using "ground-truth"
  - ❖ Answer questions using candidate summaries
  - ❖ Compare answers against "ground-truth" to evaluate opinions
- Key differences from QAEval[5]
  - ❖ Answer selection strategy to capture opinionated information
    - ➢ Input for general text summarisation – articles
      - ▪ Contain significant amount of NP and NER
    - ➢ Limited number of NP and NER in reviews
  - ❖ Answer verification strategy
    - ➢ QA use exact match or F1 score to evaluate correctness of answer
    - ➢ Does not allow abstractive answers
    - ➢ Does not consider information in question or passage

Experiments and Results

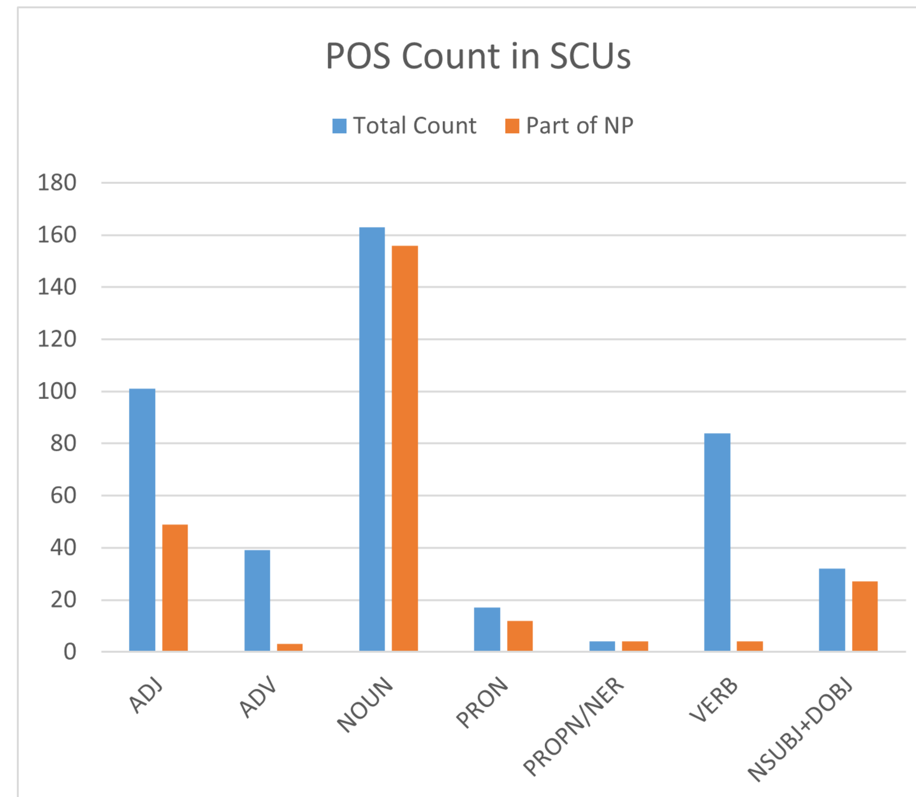

Figure 3: Part of Speech(POS) tagging of SCUs.

- Limited number of NER in reviews
- NP only captures limited:
  - ❖ ADJ
  - ❖ ADV
  - ❖ VB
- Limited NSUBJ+DOBJ:
  - ❖ Mostly exist in full sentences
  - ❖ SCU – clause not sentence

metric	pearson	spearman	kendall
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-F1	0.409	0.416	0.29
RunQA-F1	0.460	0.484	0.344
RunQA-LERC	**0.597**	**0.575**	**0.400**

Table 1: Pearson, Spearman and Kendall correlation coefficient between the metrics' scores and human annotations of **coverage/recall**.

metric	pearson	spearman	kendall
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-F1	0.555	0.555	0.409
RunQA-F1	0.551	0.654	0.475
RunQA-LERC	**0.714**	**0.712**	**0.542**

Table 2: Pearson, Spearman and Kendall correlation coefficient between the metrics' scores and human annotations of **focus/precision**.

Correlation with Human Judgement
- Human annotations collection – Amazon Mechanical Turk*:
  - ❖ Coverage/recall of information
  - ❖ Focus/precision of information
- Calculated correlation between human and metrics' scores – a good metric to be close to human judgement as possible
- RunQA-LERC best correlated with human
- Changing answer selection strategy improves performance
- Changing answer verification strategy also improves

Robustness Test
- Metric – consistently & reliably rank summaries based on quality
- 2 Systems:
  - ❖ Human – ideal
  - ❖ Copycat[7]
- Score A: Human vs. Reference
- Score B: Copycat vs. Reference
- Expected: A > B
- $Accuracy = (Number\ of\ A > B)/(Total\ Number)$
- Aim: whether a metric constantly give the human system a higher score

Metric	Accuracy
ROUGE-1	68.33%
ROUGE-2	52.78%
ROUGE-L	63.89%
BERTScore	54.44%
MoverScore	80.56%
QAEval-F1	77.78%
RunQA-F1	82.22%
RunQA-LERC	**93.33%**

Table 3: The percentage of each metric successfully assign a higher score for the ground-truth summary (human system).

- RunQA-LERC most reliable
- BERTScore and ROUGE family close to random guessing
  - ❖ Not sensitive to opinion
  - ❖ Rate summaries base on surface-level matching
- QA-based metrics good at ranking systems:
  - Evaluating fine-grained information than similarity based on text

Conclusion

Use RunQA for opinion summarisation evaluation
- Evaluate summaries by opinion consensus
- Better correlated with human judgements
- RunQA-LERC most robust

References

1. *Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).*
2. *Deutsch, D., & Roth, D. (2020). Understanding the extent to which summarization evaluation metrics measure the information quality of summaries. arXiv preprint arXiv:2010.12495.*
3. *Tay, W., Joshi, A., Zhang, X. J., Karimi, S., & Wan, S. (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 (pp. 52-60).*
4. *Novikova, J., Dusek, O., Curry, A. C., & Rieser, V. (2017, January). Why We Need New Evaluation Metrics for NLG. In EMNLP.*
5. *Deutsch, D., Bedrax-Weiss, T., & Roth, D. (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.*
6. *Chen, A., Stanovsky, G., Singh, S., & Gardner, M. (2020, November). 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) (pp. 6521-6532).*
7. *Bražinskas, A., Lapata, M., & Titov, I. (2020, July). Unsupervised Opinion Summarization as Copycat-Review Generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5151-5169).*