Instructions for ACL 2019 Proceedings

Anonymous ACL submission

Abstract

Automatic summarization research has made substantial progress thanks to novel methods and datasets. Despite this progress, most manual evaluation focuses on fluency but not on the content of the summaries. Those that do, typically use a single reference summary for comparison, either directly or through questions, which introduces a bias towards a single correct answer for a task where this assumption doesn't hold. To address this issue, we propose Highlight-bAsed Evaluation of Single document Summarization (HArnESS) that provides assessment of a summary against the original document, facilitated through manually highlighted salient content. The highlights lower the variability of the judges' assessment and are reusable in future studies. Furthermore it does not require expert annotators, avoids reference bias and provides absolute instead of ranked evaluation of the systems.

1 Introduction

Research in automatic summarization has made headway over the years with single document summarization as the front-runner due to the availability of large datasets (Sandhaus, 2008; Hermann et al., 2015; Narayan et al., 2018a) and novel methods (See et al., 2017; Kryściński et al., 2018; Narayan et al., 2018a). Despite this progress, there is a large number of studies (e.g., See2017) that didn't perform any manual evaluation at all. Those that do, typically did not take account of bias that comes from comparing to a single reference summary, either directly (Celikyilmaz et al., 2018; Tan et al., 2017) or through questions (Narayan et al., 2018a,b); or have little agreements on how to evaluate content of the summaries. As the rest of them that don't, they were mostly relying on automatic evaluation such as ROUGE (Lin, 2004), but it has been shown in Louis and Nenkova (2013) that for single document summarization case ROUGE is inaccurate due to reference bias.

Up to now, too little attention has been paid to build a consensus in manual evaluation framework which has eventually lead to a fragmented field and many unnecessary system evaluation rerun due to non-reusable results. Summarization field used to have a common manual evaluation practice that is established through DUC such as Pyramid method (Nenkova and Passonneau, 2004) but then the method would need expert annotators for every new dataset which is costly. Furthermore, the method is also prone to reference bias when there is only one summary available for every document.

In this paper we propose Highlight-bAsed Evaluation of Single document Summarization (HArnESS) which makes three contributions: (1) We present a new way to assess a content of summary by comparing it against the original document, facilitated through manually highlighted salient document. (2) We provide a complete manual evaluation framework for evaluating the quality of summary that yields reusable result and does not require expert annotators. (3) We provide a new augmented annotation for XSum dataset.

2 Data

For our experiment, we use the extreme summarization (XSum) Dataset (Narayan et al., 2018a) which comprises of BBC article and summary pairs. The summary in the XSum dataset has a big proportion of novel unigrams compared to other more popular dataset such as CNN/Daily Mail or NY Times. This makes the dataset suitable for our experiment since we don't want the judges to be biased towards extractive methods which makes the process of assessment easier compared to the abstractive methods. In other words, we expect in

an abstractive setting, the task of reading an article will be more laborous and prone to disagreement between judges.

We didn't use the whole test set portion for evaluation, but only sampled 50 articles from it for the annotation and evaluation purpose following the human evaluation practice from the Narayan et al. (2018a)'s author.

For the peer summaries, we use the Topic-aware Convolutional Sequence to Sequence (T-CONVS2s) model [1] (Narayan et al., 2018a) and Pointer-Generator Networks model [2] (See et al., 2017) as the comparisons. We obtained both summary result by directly running the code on our sampled dataset.

3 Highlight Annotation

The highlight annotation is a manual process where we ask human to read the XSum article and then highlight words or phrases that are considered as important for the article. To perform the task, we use the Amazon HIT (Human Intelligence Task). Each of the 50 articles is annotated ten times done separately.

We gave the Turkers a guidance on how to do the highlight task using the 5W1H (Who, What, When, Where, Why and How) principles (Robertson, 1946) that is a common practice in journalism. Since the dataset is a news article, the most important information would be part of the articles that answer the 5W1H question. However, we don't expect every single annotation to cover all the answer to the 5W1H, instead we want the annotation to have a more diverse highlights which cover different viewpoints of the article which is why we limited each annotator to a maximum of 30 words highlight.

At the end of the task, we performed a sanity checking to ensure the reliability of the annotation result by asking the Turker to answer a True/False question about part of the article. We expect each Turker that had performed the highlight would be able to answer the question correctly since they had read the whole article. We rejected all annotations that failed to correctly answer the sanity question. Table 1 shows three out of ten annotation for one article [3].

[1] https://github.com/EdinburghNLP/XSum

[2] https://github.com/abisee/pointer-generator

[3] https://www.bbc.co.uk/newsround/17546358

No.	Highlighted Words
1	['Mr Bezos', 'found the engines from the Apollo 11 space rocket', '4,300 metres below the surface of the Atlantic Ocean', 'going to ask Nasa', 'display', 'engines', 'Museum of Flight']
2	['engines broke off from the spaceship after blast off', 'crashed somewhere in the Atlantic Ocean .', 'Mr Bezos', 'found the engines from the Apollo 11']
3	['engines from the Apollo 11 space rocket - the craft that carried the first men to the moon in 1969 .', 'Mr Bezos', 'Atlantic Ocean', 'Apollo 11 unfold on television']

Table 1: Three samples of highlight annotation for article ID 17546358

4 Evaluation

We evaluated two models: TConvs2s (Narayan et al., 2018a) and PTGen (See et al., 2017) using the following evaluation frameworks. For the evaluation, we use the direct assessment approach instead of relative ranking as this would enable us to reuse the results for subsequent evaluation with different models furthermore we isolate each model summary assessment in its own task thereby avoiding a bias from having seen different models. In addition to that, we also use a 1-100 rating scale instead of 1-5 rating scale that is normally used for summarization approach. For each model's summary, we ask three different Turkers to judge the quality.

4.1 Content Evaluation

In content evaluation, we run two different settings: the assessment by comparing the summary against the reference summary and against the article directly.

Comparison against the reference summary The head-to-head comparison of model and reference summary is a norm in DUC [4] and automatic assessment such as ROUGE and BLEU. For this task we ask the Turker to assess a randomly model summary against the reference summary of the same article using the recall and precision metrics.

The recall and precision are represented by these two statements: *'All important information is present in the summary'* and *'Only important information is in the summary.'* The first statement denotes how good is the article coverage of the

[4] Document Understanding Conference http://duc.nist.gov/use

PTGEN			
	precision	recall	f1
mean	44.24	38.34	35.83
std	27.08	28.20	25.57
$\hat{cv}$	0.78	0.84	0.82
TCONVS2S			
	precision	recall	f1
mean	46.75	36.45	36.83
std	25.86	27.91	25.19
$\hat{cv}$	0.78	0.90	0.85

Table 2: The mean, standard deviation and coefficient of variability for PTGEN and TCONVS2S models using comparison against reference summary.

summary, while the second one denotes how precise is the information conveyed by the summary. The benefit of having an analogous metric to the automatic assessment would means that we are able to calculate the F_1 score from the recall and precision just like the automatic score. We also measure the variability of the scores by calculating the coefficient of variation (cv). The cv score is the ratio between the sample standard deviation and sample mean as the following equation.

$$c_v = \frac{\sigma}{\bar{x}} \quad (1)$$

where σ is the standard deviation, and $\bar{x}$ is the mean. Since, our samples are quite small, we need to use the unbiased version as the following equation.

$$\hat{c_v} = (1 + \frac{1}{4n})c_v \quad (2)$$

Using the coefficient of variation we can compare the variability of different models.

Table 2 shows the evaluation result for PTGEN and TCONVS2S models.

Comparison against the article For the task, we gave the Turker one summary (could be model or reference summary) and the related article. There are however, two different configurations: facilitated with highlight guidance (guided) and without any facilitation (unguided). The highlight guidance is intended to help the Turker to decide the summary content quality when making the judgment. We use the same rating scale of 1-100 and the same metrics: precision and recall. Table 3 shows the evaluation result for Reference, PTGEN and TCONVS2S models.

4.2 Sentence Quality Evaluation

In sentence quality evaluation, we ask the Turker to rate the sentence summary alone on two metrics: sentence clarity and fluency. For clarity, the

Guided Reference				Unguided Reference		
	Prec	Rec	F1	Prec	Rec	F1
mean	67.90	56.83	55.25	66.01	52.45	50.67
std	25.36	26.26	25.43	26.36	29.18	25.28
$\hat{cv}$	0.49	0.63	0.61	0.48	0.67	0.61
Guided PTGEN				Unguided PTGEN		
	Prec	Rec	F1	Prec	Rec	F1
mean	50.94	44.41	42.17	48.57	39.21	37.80
std	27.54	30.68	26.28	27.31	30.11	26.84
$\hat{cv}$	0.73	0.86	0.81	0.73	0.90	0.84
Guided TCONVS2S				Unguided TCONVS2S		
	Prec	Rec	F1	Prec	Rec	F1
mean	57.42	49.95	47.00	52.55	41.04	39.25
std	29.63	31.23	26.91	27.45	26.18	24.38
$\hat{cv}$	0.67	0.80	0.71	0.75	0.83	0.81

Table 3: The mean, standard deviation and coefficient of variability for Guided and unguided Reference, PTGEN and TCONVS2S models using comparison against article.

Turker is asked whether the summary is easy to be understood of which there should be no difficulties in identifying the referents of the noun phrases (every noun/place/event should be well-specified) or understanding the meaning of the sentence. While for fluency, the Turker is asked whether the summary sounds natural and has no grammar-related problem that makes the text difficult to read.

5 Discussion

In this section, we are going to discuss several insights gained from the annotation and evaluation processes.

5.1 Annotation Agreement

The average inter-annotator's agreement for 50 articles measured using Fleiss Kappa (Fleiss, 1971) is 0.185. Table 1 shows the inter-annotator agreement for 50 articles. This result shows only a slight agreement for each article, however, we considered it as normal since one article may have different viewpoints, for example in Table 1 there are three different viewpoints on one article.

5.2 Variability Analysis

The coefficient of variability results show a slight improvement for the guided models compares to the unguided ones which means that the judges are more agreeing with each other over the perceived quality of a summary. For the PTGEN model, the guided F_1 score is 0.03 lower than the unguided while the for the TCONVS2S model, the guided F_1 score is 0.1 lower than the unguided. Meanwhile there is no difference between the guided

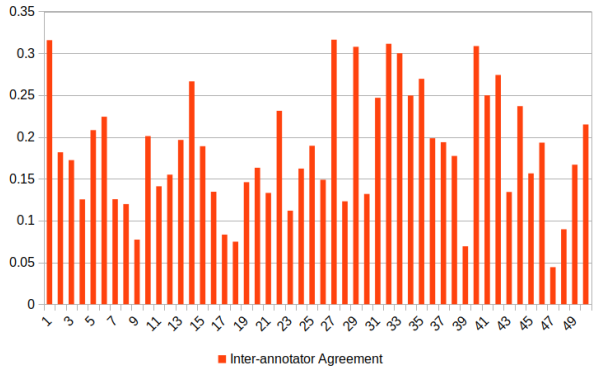

Figure 1: The inter-annotator agreement of highlight annotation for 50 articles

and unguided. To understand the cause, we look into several examples to see why this has happened.

The first example is the article ID 36328049[5]. Table 4 and Table 5 shows the result and summary text for the models and reference.

TCONVS2S				
F1 Guided	F1 Unguided	Guided-Unguided	cv Guided	cv Unguided
1.98	64.56	-62.58	0	0.44
PTGEN				
F1 Guided	F1 Unguided	Guided-Unguided	cv Guided	cv Unguided
67.61	41.32	26.29	0.37	1.05

Table 4: The result from PTGEN, TCONVS2S, and Reference for Article ID 36328049.

Model	Summary Text
PTGEN	a nigerian refugee has been rescued from a number of pupils after being kidnapped by islamic state (is) militants in northern nigeria .
TCONVS2S	nigerian schoolgirl who was abducted by boko haram militants in nigeria has been reunited with her parents .
REFERENCE	the first of the missing nigerian schoolgirls to be rescued since her capture two years ago has had an emotional reunion with her mother .

Table 5: The summaries text from PTGEN, TCONVS2S, and Reference for Article ID 36328049.

For both models, the cv of the unguided is higher than the guided ones. We checked closely into the highlights and the articles (see Appendix), and we found that the cause of the high variance is because without the guidance, judges are more to disagree the salient contents since the article is long and has several viewpoints. But with proper guidance, judges are more consistent in giving a rating.

Another example is article ID 38204334[6]. In this article, the TCONVS2s successfully shows the most important part of the article which are also highlighted which is why the judges are more consistent in giving a high rating.

Model	Summary Text
PTGEN	storm desmond battered parts of the uk and lancashire last winter , according to new figures from the office for national statistics (ceh)
TCONVS2S	parts of the uk have been battered by flooding in the past year , according to a new report .
REFERENCE	flooding across parts of the uk last winter was the most extreme on record , experts have said .

Table 6: The summaries text from PTGEN, TCONVS2S, and Reference for Article ID 38204334.

TCONVS2S				
F1 Guided	F1 Unguided	Guided-Unguided	cv Guided	cv Unguided
86.76	37.89	48.87	0.20	0.34
PTGEN				
F1 Guided	F1 Unguided	Guided-Unguided	cv Guided	cv Unguided
41.18	62.02	-20.84	0.67	0.70

Table 7: The result from PTGEN, TCONVS2S, and Reference for Article ID 38204334.

6 Related Works

Based on ranking scheme there are three approaches: rating scale (Likert, 1932) used by Kryściński et al. (2018), paired comparison (Thurstone, 1994) used by Fan et al. (2018); Celikyilmaz et al. (2018); and best-worst scaling (Woodworth and G, 1991) used by Narayan et al. (2018a). Then based on subject of comparison, there are head-to-head comparison between system and reference summary (Celikyilmaz et al., 2018) and ground truth comparison using the document directly (Narayan et al., 2018a; Kryściński et al., 2018). Finally there are also different method in evaluating, for example question-answering method (Clarke and Lapata, 2010; Narayan et al., 2018a). In addition to that, several researches (Nallapati et al., 2016; See et al., 2017; Gehrmann et al., 2018) did not employ

[5] https://www.bbc.co.uk/news/world-africa-36328049

[6] http://www.bbc.co.uk/news/uk-38204334

manual evaluation but instead opted to do qualitative analysis directly on the system summary.

Acknowledgments

The acknowledgments should go immediately before the references. Do not number the acknowledgments section. Do not include this section when submitting your paper for review.

Preparing References:

Include your own bib file like this:

```
\bibliographystyle{acl_natbib}
\bibliography{acl2019}
```

where `acl2019` corresponds to a acl2019.bib file.

References

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep Communicating Agents for Abstractive Summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

James Clarke and Mirella Lapata. 2010. Discourse constraints for document compression. *Computational Linguistics*, 36(3):411–441.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable Abstractive Summarization. In *ACL 2018 Workshop on Neural Machine Translation and Generation*.

Joseph L. Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, 76(5):378–382.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-Up Abstractive Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Neural Information Processing Systems*, pages 1–14.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving Abstraction in Text Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Proceedings of the workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004*, (1):25–26.

Annie Louis and Ani Nenkova. 2013. Automatically Assessing Machine Summary Content Without a Gold Standard. *Computational Linguistics*, 39(2):267–300.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. *Proceedings of CoNLL*, pages 280–290.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018a. Don't Give Me the Details, Just the Summary! Topic-aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. Ranking Sentences for Extractive Summarization with Reinforcement Learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1747–1759, Stroudsburg, PA, USA. Association for Computational Linguistics.

A Nenkova and R Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. *Proceedings of HLT-NAACL*, 2004:145–152.

D. W. Robertson. 1946. A Note on the Classical Origin of " Circumstances " in the Medieval Confessional. *Studies in Philology*, 43(1):6–14.

Evan Sandhaus. 2008. The New York Times Annotated Corpus.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the ACL*.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive Document Summarization with a Graph-Based Attentional Neural Model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1171–1181, Stroudsburg, PA, USA. Association for Computational Linguistics.

L. L. Thurstone. 1994. A Law of Comparative Judgment. *Psychological review*, 101(2):255–270.

Jordan J Louviere Woodworth and George G. 1991. Best-worst scaling: A model for the largest difference judgments. *University of Alberta: Working Paper*.

A Appendices

Appendices are material that can be read, and include lemmas, formulas, proofs, and tables that are not critical to the reading and understanding of the paper. Appendices should be **uploaded as supplementary material** when submitting the paper for review. Upon acceptance, the appendices come after the references, as shown here. Use `\appendix` before any appendix section to switch the section numbering over to letters.

B Supplemental Material

Submissions may include non-readable supplementary material used in the work and described in the paper. Any accompanying software and/or data should include licenses and documentation of research review as appropriate. Supplementary material may report preprocessing decisions, model parameters, and other details necessary for the replication of the experiments reported in the paper. Seemingly small preprocessing decisions can sometimes make a large difference in performance, so it is crucial to record such decisions to precisely characterize state-of-the-art methods.

Nonetheless, supplementary material should be supplementary (rather than central) to the paper. **Submissions that misuse the supplementary material may be rejected without review.** Supplementary material may include explanations or details of proofs or derivations that do not fit into the paper, lists of features or feature templates, sample inputs and outputs for a system, pseudo-code or source code, and data. (Source code and data should be separate uploads, rather than part of the paper).

The paper should not rely on the supplementary material: while the paper may refer to and cite the supplementary material and the supplementary material will be available to the reviewers, they will not be asked to review the supplementary material.