Human Factors in Software and Systems Engineering, Vol. 94, 2023, 22–27
https://doi.org/10.54941/ahfe1003766

The Removal of Irrelevant Human Factors in a Multi-Review Corpus Through Text Filtering

Aaron Moody, Makenzie Spurling, and Chenyi Hu

University of Central Arkansas, Conway, AR 72035, USA

ABSTRACT

Generating a high-quality explainable summary of a multi-review corpus can help people save time in reading the reviews. With natural language processing and text clustering, people have generated both abstractive and extractive summaries on a corpus consisting of 967 reviews about a refurbished phone on Amazon (Moody et al. 2022). However, the overall quality of the summaries needs further improvement. Noticing the online reviews in the corpus come from a diverse population, we take an approach of removing irrelevant human factors through pre-processing. Apply available pre-trained models together with reference based and reference free metrics, we filter out noise in each review automatically prior to summary generation. Our computational experiments evident that one may significantly improve the overall quality of an explainable summary from such a pre-processed corpus than from the original one. It is suggested of applying available high-quality pre-trained tools to filter noises rather than start from scratch. Although this work is on the specific multi-review corpus, the methods and conclusions should be helpful for generating summaries for other multi-review corpora.

Keywords: Multi-review corpus, Natural language processing, Text summarization, Text filtering

INTRODUCTION

In this study, we use the corpus in (Moody et al. 2022) which consists of 967 reviews on a refurbished cellphone on Amazon. Each review consists of a five-star scaled numeric ranking together with text comments. So, we denote a review from a reviewer i as a tuple $r_i = (s_i, d_i)$. In which, s_i and d_i represent the numeric ranking and the document containing text comments, respectively. Table 1 below lists the number of reviews in each of the five-star rankings.

From the table, we have the average star ranking 3.9, which is different from Amazon's overall ranking 4.4. "Amazon calculates a product's star rating using machine-learned models instead of a simple average." --Amazon. A human user is expected to read these 967 text comments manually to comprehend the 4.4 ranking by Amazon. Reading these 967 reviews manually online is a tedious and unfeasible task. Most of the time, people avoid doing so by reading only a small portion of reviews. However, by doing this, the

Table 1. Star ranking frequencies of the reviews.

Ranking	1-star	2-star	3-star	4-star	5-star
Count	179	48	57	119	564

Table 2. ROUGE scores of extractive summaries of each star-ranking cluster.

Ext. summary	1-star	2-star	3-star	4-star	5-star
R-1 recall	1.0	0.8333	1.0	0.8571	1.0
R-1 precision	0.1114	0.1562	0.2120	0.1389	0.0580
R-1 F score	0.1985	0.2250	0.3415	0.2233	0.1086
R-2 recall	0.9790	0.8130	0.9754	0.8310	0.9776
R-2 precision	0.0663	0.1001	0.1476	0.0985	0.0271
R-2 F score	0.1227	0.1731	0.2487	0.1619	0.0524
R-*l* recall	1.0	0.8333	1.0	0.8571	1.0
R-*l* precision	0.1114	0.1562	0.2120	0.1389	0.0580
R-*l* F score	0.1985	0.2250	0.3415	0.2233	0.1086

user only gets a snapshot of the performance of a product rather than what is holistically true. By applying text summarization in natural language processing, people can generate a summary from the reviews for human to read. However, such summaries in practice may suffer from two issues. They are either missing valuable information or are hard for humans to read. These issues are because of that a multi-review corpus usually consists of texts from a diverse population with various background and language style. In other words, there are noises due to human factors in such corpora and we should eliminate irrelevant text within a multi-review corpus first. Prior to our discussion, let us briefly review some previous results.

A BRIEF REVIEW OF RELATED PREVIOUS WORK

Applying text clustering and text summarization, Moody et al. (2022) proposed the algorithm below for an explainable summary.

Algorithm 1: Obtain a summary of a multi-review corpus cluster-wise.
Input: $D = \{(s_i, d_i)\}$, a set of eligible reviews
Output: A summary of D
$\{D_j\} \leftarrow$ clustering D
For each D_j
  $S_j \leftarrow$ text summary of D_j
Return $\{\mathbf{S_j}\}$

In that work, a corpus of reviews is clustered according to star ranking. Then, reviews in each cluster are summarized. Table 2 gives the ROUGE 1, 2, and *l* scores (Lin, 2004) of extractive summaries of each cluster.

Because of the summaries are formed with important sentences extracted from each cluster directly, the recalls are near 1. However, the precision and F-1 scores are low. In that work, a hierarchical graph attention (HGAT) network (Zhan et al. 2021) has also been applied to generate abstractive summaries. But these summaries are not easy to read though the ROUGE scores are improved.

FILTERING OUT NOISES WITH PRE-TRAINED TRANSFORMERS

The overall quality of the above summaries needs improvement. Before doing so, we should identify potential causes first. Examining the corpus again, we find that the 967 reviews are in various lengths and writing styles. This is because each reviewer comes from a diverse population. This explains why some reviews contain excessive text that is irrelevant to the user's star-rankings. Moreover, a reference is needed in calculating ROUGE scores. However, the original corpus contains redundant information and noises, which we are trying to remove. Using the original corpus as the reference in the previous work is not reasonable. Therefore, we should filter out irrelevant noises caused by human factors first before summarizing the corpus.

Building Filters With Pre-Trained Transformers

Filtering irrelevant noises from multi-reviews itself is a challenging task in machine learning and AI. Advanced deep neural network architectures like transformers (Vaswani, A., et al. 2017) have brought significant improvement to this task. However, training transformers requires high computational capacities together with large volumes of labeled training data. Instead of training for specific tasks, publicly available pre-trained models have become the trend very recently. For example, Google's Bidirectional Encoder Representations from Transformers (BERT, Devlin et al. 2018 and Tenney et al, 2019) and Generative Pre-trained Transformer (GPT) of OpenAI have changed the landscape of NLP. In addition, Hugging Face (huggingface.co/) hosts hundreds of pre-trained transformers available for text summarization. Following the trend, we apply four models from Hugging Face to build our filters. They are D-Pegasus[1] (Shleifer at al. 2010 and Zhang et al. 2019), two BART (Lewis et al. 2019) models BART-Lid[2] and FT-BART[3] and a T5 (Text-to-Text Transfer Transformer, Raffel et al. 2019) model Titlewave-T5[4].

Algorithm 2: Pre-process a multi-review corpus through item summarization.
Input: a corpus $D = \{d_i\}$
Output: a pre-processed corpus A obtained from D
for $d_i \in D$ **do**
  $p \leftarrow$ D-Pegasus(d_i) # p is the summary with distil-pegasus
  b $\leftarrow$ BART-Lid(d_i) # b is the summary with Lydia-BART
  $t \leftarrow$ Titlewave-T5(d_i) # t is the summary with titlewave-t5
  $f \leftarrow$ FT-BART(d_i) # f is summary with finetuned-BART
  $a_i \leftarrow max\,(p, b, t, f)$ # picking the one with max average ROUGE F-scores
end for
  $A \leftarrow \{a_i\}$
return A

[1] huggingface.co/sshleifer/distill-pegasus-xsum-16-8
[2] huggingface.co/lidiya/bart-large-xsum-samsum
[3] huggingface.co/knkarthick/bart-large-xsum-samsum
[4] huggingface.co/tennessejoyce/titlewave-t5-base

In the algorithm above, we use the maximum average ROUGE F-score (R1-F + R2-F + R3-F)/3, which is reference based, as the selection criterion. As an alternative, we also applied a reference-free evaluation metric, Shannon score (Egan et al. 2021) as a selection criterion. In addition, the Flesch reading ease metric is used as a metric for readability.

Beyond summarizing each review abstractively, we use another algorithm to pre-process the original corpus through breaking it into sentences. Each sentence is then treated as a summary of the corpus. We evaluate how well each sentence represents the corpus with the metrics mentioned above. The top K scoring sentences form a pre-processed corpus. The idea here is that the most relevant sentences will have the highest evaluation scores, and sentences with lower scores will contain irrelevant information that we want to exclude from the corpus.

Human Intelligence Involvement

Using one or a combination of the selection criteria above with Algorithm 2, we can obtain a pre-processed corpus A from the original corpus D automatically. We refer to A as an abstractive corpus because each review is summarized using an abstractive-based transformer model first before noise removal. In this project, we also form a humanized corpus through preprocessing each review manually, which is very tedious from scratch. Instead, for practicality, we apply human intelligence on the *p, b, t, f* in Algorithm 2 to make the selection. The three corpora: abstractive, humanized, and original are used in our computational experiments to derive summaries that explain the multi-review corpus.

RESULTS OF COMPUTATIONAL EXPERIMENTS

To examine the effects of removing irrelevant human factors from a multi-review corpus, we run several experiments computationally. Table 3 below compares the quality of cluster-wise abstractive summaries from the three corpora. The table compares the quality of summaries derived from each star-ranking cluster with various methods (including BERT). We use ROUGE scores as the metric. Here are a few highlighted observations from Table 3.

1. Observing each row of the table, we can find that the quality of abstractive summary derived from the original corpus is much worse than that from the two corpora with noise removal. This implies that filtering irrelated human factors can improve the quality of abstractive summary. And for a fair comparison, we used the original corpus as the reference when calculating ROUGE scores for the summaries. This shows that a reduction of corpus size impacts the ROUGE scores of corpus summary positively.
2. The ROUGE scores of summaries derived from abstractive corpus and humanized corpus are mixed. In some cases, one is higher than the other. It is just opposite in other cases. However, the differences are not significant. This suggests that our automated filtering process is compatible with human involved selection process. In other words, it is worthwhile

Table 3. Comparisons of ROUGE scores on summaries from three corpora*.

Star	Method	Abstractive Corpus			Humanized Corpus			Original Corpus		
		R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
1	D-Pegasus	1.21	0.34	1.21	2.20	0.79	2.20	0.51	0.12	0.51
	BART-Lid	3.99	1.31	3.99	5.72	0.78	5.36	1.70	0.34	1.70
	Titlew.-T5	1.61	0.41	1.61	2.91	0.78	2.91	0.68	0.14	0.68
	FT-BART	3.99	1.31	3.99	5.72	0.78	5.36	1.17	0.34	1.70
	BERT	27.4	15.7	27.4	25.6	8.97	23.9	11.8	4.65	11.5
2	D-Pegasus	4.06	1.23	4.06	9.24	2.42	21.7	1.73	0.36	1.72
	BART-Lid	25.6	15.9	25.6	22.9	6.85	21.7	10.7	4.03	10.1
	Titlew.-T5	3.71	1.23	3.33	6.77	2.43	6.77	1.44	0.31	1.44
	FT-BART	16.2	8.40	16.2	18.05	4.29	17.32	6.58	1.91	6.44
	BERT	35.3	26.0	35.3	27.8	6.62	26.0	14.5	6.15	14.2
3	D-Pegasus	4.47	1.81	4.47	7.35	2.81	7.23	1.72	0.43	1.72
	BART-Lid	16.7	6.97	16.7	16.7	3.30	16.1	6.71	1.66	6.57
	Titlew.-T5	2.26	0.66	2.26	3.75	0.81	3.75	0.86	0.10	0.86
	FT-BART	13.5	6.84	13.5	13.0	4.08	13.3	5.35	1.55	5.21
	BERT	29.6	19.6	29.6	33.0	13.2	31.9	11.2	4.57	10.9
4	D-Pegasus	2.71	0.91	2.71	4.97	1.23	4.97	1.04	0.31	1.04
	BART-Lid	12.4	5.83	12.4	19.1	5.38	17.1	4.79	1.54	4.79
	Titlew.-T5	1.81	0.50	1.81	3.19	0.62	3.19	0.70	0.15	0.70
	FT-BART	7.07	2.56	7.07	9.52	2.23	9.09	2.68	0.65	2.68
	BERT	26.5	15.1	26.5	26.7	7.39	25.9	10.2	3.45	10.2
5	D-Pegasus	1.72	0.41	1.72	2.55	0.64	2.55	0.69	0.12	0.69
	BART-Lid	3.97	1.29	3.97	6.09	1.81	5.91	1.51	0.40	1.51
	Titlew.-T5	1.38	0.28	1.38	2.01	0.53	0.52	0.52	0.09	0.52
	FT-BART	2.74	0.85	2.74	4.52	1.23	4.34	1.08	0.26	1.08
	BERT	30.1	16.6	30.1	33.0	13.6	31.5	12.0	4.82	11.1

* ROUGE scores in the table are multiplied by a hundred for easier viewing.

to fine-tune the AI approach rather than spending precious human power to filter out irrelevant text in multi-reviews for an explainable summary.

3. Comparing ROUGE scores in each cluster on all three corpora, we find that the results obtained with BERT are significantly better than that with D-Pegasus, BART-Lid, FT-BART, and Titlewave-T5. This is because of that BERT has been well trained with very large volumes of data with powerful TPUs at Google. Due to the significant resource requirement of training an AI platform for NLP, using a trustworthy well pre-trained system should be a very good choice in practice.

Similar results are observed in our experiments on generating extractive summary and abstractive-extractive mixed summarization.

CONCLUSION

Computationally generated explainable summaries of multi-review corpora can help human users to comprehend the reviews for decision making efficiently and effectively. The study on this specific dataset suggests that prior to summarize a multi-review corpus one should pre-process it first to filter

out noises due to irrelevant human factors. In addition, this can be done with pre-trained machine learning models. Both reference based and reference free metrics, such as ROUGE score and Shannon score, can be applied as selection criteria. Instead of generating a filter from scratch for a specific multi-review corpus, one should consider a well-trained NLP platform first with possible fine-tune for further quality improvement.

OpenAI released ChatGPT (GPT 3.5) on November 30, 2022, and GPT 4 on March 14, 2023. We are currently working on utilizing these newly available tools to filter out noises and to generate explainable summaries for the multi-review corpus.

ACKNOWLEDGMENT

This work is partially supported by the US National Science Foundation through the grant award NSF/OIA-1946391.

REFERENCES

Devlin, J., Chang, M.-W., Lee, K., Toutanova (2018): BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805.

Egan, N., et al. (2021). Play the Shannon Game with Language Models: A Human-Free Approach to Summary Evaluation. arXiv:2103.10918.

Lewis, M., and *et al* (2019).: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. https://arxiv.org/abs/1910.13461

Lin, C-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Moody, A., Hu, C., Zhan, H., Spurling, M., Sheng, V. S. (2022). Towards explainable summary of crowdsourced reviews through text mining. In. Communications in Computer and Information Science, vol 1601 (pp. 528–541). Springer, Cham.

Raffel, C., and *et al*. (2019): Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, https://arxiv.org/abs/1910.10683.

Shleifer, S., Ruth, A. M. (2010): Pre-trained Summarization Distillation, arXiv:2010.13002.

Tenney, I., Das, D., and Pavlick, E. (2019): BERT Rediscovers the Classical NLP Pipeline, https://arxiv.org/abs/1905.05950.

Vaswani, A., et al (2017): Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

Zhan, H., Zhang K., Hu, C., Sheng VS. (2021): HGATs: Hierarchical Graph Attention Networks for Multiple Comments Integration, Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. (2019): PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization, https://arxiv.org/abs/1912.08777.