Towards Standardized Manual Evaluation for Single Document Automatic Summarization

Anonymous NAACL submission

Abstract

Having a good summary evaluation approach is important for automatic summarization field, however, in automatic summarization community there is still no consensus in manual evaluation practice. This difficult situation is coupled with the release of recent large newswire dataset for single document summarization that are characterized by only having a single reference summary per document raise an issue of reference bias for common practice in automatic evaluation such as ROUGE. In this paper, we present Summ-Eval, a manual evaluation toolkit and practice for single document summarization which: (1) provides a means of getting comparable and reusable results, (2) provides evaluation criteria and validation scheme, (3) resolves the reference bias, and (4) is accessible for non-expert evaluators

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. Many single document summarization researches (See et al., 2017; Kryściński et al., 2018; Narayan et al., 2018) are aiming to produce the highest quality of summary indicated by its capability to capture all the important content of the document while still maintain sentence fluency and clarity. Hence, having a good summary evaluation approach is important for the field as it provides a means to rank different systems.

ROUGE (Lin, 2004) for automatic evaluation has seen a wide adoption, but recent releases of newswire dataset such as NY Times (Sandhaus, 2008), CNN/Daily Mail (Hermann et al., 2015), and XSum (Narayan et al., 2018) are characterized by only having a single reference summary per document which makes ROUGE becomes less accurate due to reference bias (Louis and Nenkova, 2013). As such, to get a reliable assessment a manual evaluation is needed. However, there is still no consensus in manual evaluation practice yet in automatic summarization field, whereas the machine translation have already adopted practice and toolkit for manual evaluation like Appraise (Federmann, 2012). Consequently, many researches report manual evaluation results that are often non-reusable for subsequent researches due to the lack of transparency (e.g. evaluation prompts and setup), and the inherent nature of the evaluation process (e.g. paired comparison and best-worst scaling). This cause many researches have to keep repeating evaluating the same systems which is a laborious and costly task. Other prevailing issues are differing views on evaluation criteria and validation scheme, reference bias, and access to expert evaluators.

To show how different each of the manual evaluation, we look into recent practices in a single document summarization field. We divided all these approaches into three category: ranking scheme, subject of comparison, and method.

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. (2018). 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., 2018; Kryściński et al., 2018). Finally there are also different method in evaluating, for example question-answering method (Clarke and Lapata, 2010; Narayan et al., 2018). In addition to that, some researches (Nallapati et al., 2016; See et al., 2017; Gehrmann et al., 2018) did not employ manual evaluation but in-

stead opt to do qualitative analysis directly on the system summary.

To resolve all the aforementioned issues, we present a toolkit as well as a standardized practice for manual evaluation in single document summarization, *Summ-Eval*: a manual evaluation toolkit and practice which: (1) provides a means of getting comparable and reusable results, (2) provide evaluation criteria and validation scheme, (3) resolves the reference bias, and (4) is accessible for non-expert evaluators.

2 Related Works

Test

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.

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*.

Christian Federmann. 2012. Appraise: an Open-Source Toolkit for Manual Evaluation of MT Output. *The Prague Bulletin of Mathematical Linguistics*, 98(-1):130–134.

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. 2018. 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*.

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*.

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*.