SUMMARY EXPLORER
Visualizing the State of the Art in Text Summarization

Shahbaz Syed [†*] **Tariq Yousef** [†*] **Khalid Al-Khatib** [†]

Stefan Jänicke [‡] **Martin Potthast** [†]

[†]Leipzig University [‡]University of Southern Denmark

<shahbaz.syed@uni-leipzig.de> <tariq.yousef@uni-leipzig.de>

Abstract

This paper introduces SUMMARY EXPLORER, a new tool to support the manual inspection of text summarization systems by compiling the outputs of 55 state-of-the-art single document summarization approaches on three benchmark datasets, and visually exploring them during a qualitative assessment. The underlying design of the tool considers three well-known summary quality criteria (coverage, faithfulness, and position bias), encapsulated in a guided assessment based on tailored visualizations. The tool complements existing approaches for locally debugging summarization models and improves upon them. The tool is available at https://tldr.webis.de/.

1 Introduction

Automatic text summarization is the task of generating a summary of a long text by condensing it to its most important parts. This longstanding task originated in automatically creating abstracts for scientific documents (Luhn, 1958), and later extended to documents such as web pages (Salton et al., 1994) and news articles (Wasson, 1998).

There are two paradigms of automatic summarization: *extractive* and *abstractive*. The former extracts important information from the to-be-summarized text, while the latter additionally involves paraphrasing, sentence-fusion, and natural language generation to create fluent summaries. Neural summarization approaches trained on large-scale datasets have significantly advanced both paradigms by improving the overall document understanding and text generation capabilities of the models to generate fluent summaries.

Currently, the progress in text summarization is tracked primarily using *automatic evaluation* with ROUGE (Lin, 2004) as the de facto standard for quantitative evaluation. ROUGE has proven effective for evaluating extractive systems, measuring the overlap of word n-grams between a generated summary and a reference summary (ground truth). Still, it only provides an approximation of a model's capability to generate summaries that are lexically similar to the ground truth. Moreover, ROUGE is unsuitable for evaluating abstractive summarization systems, mainly due to its inadequacy in capturing all semantically equivalent variants of the reference (Ng and Abrecht, 2015; Kryscinski et al., 2019; Fabbri et al., 2021). Besides, a reliable automatic evaluation of a summary is challenging (Lloret et al., 2018) and strongly dependent on its purpose(Jones et al., 1999).

A robust method to analyze the effectiveness of summarization models is to manually inspect their outputs from individual perspectives such as coverage of key concepts and linguistic quality. However, manual inspection requires obtaining the outputs of certain models, delineating a guideline that comprises particular assessment criteria, and ideally utilizing proper visualization techniques to examine the outputs efficiently.

To this end, we present SUMMARY EXPLORER (Figure 1), an online interactive visualization tool that assists humans (researchers, experts, and crowds) to inspect the outputs of text summarization models in a guided fashion. Specifically, we compile and host the outputs of several state-of-the-art models (currently 55) dedicated to English single-document summarization. These outputs cover three benchmark summarization datasets comprising semi-extractive to highly abstractive ground truth summaries. The tool facilitates a *guided* visual analysis of three important summary quality criteria: *coverage*, *faithfulness*, and *position bias*, where tailored visualizations for each criterion streamline both absolute and relative manual evaluation of summaries. Overall, our use cases (see Section 5) demonstrate the ability of SUMMARY EXPLORER to provide a comparative exploration of

* Equal contribution.

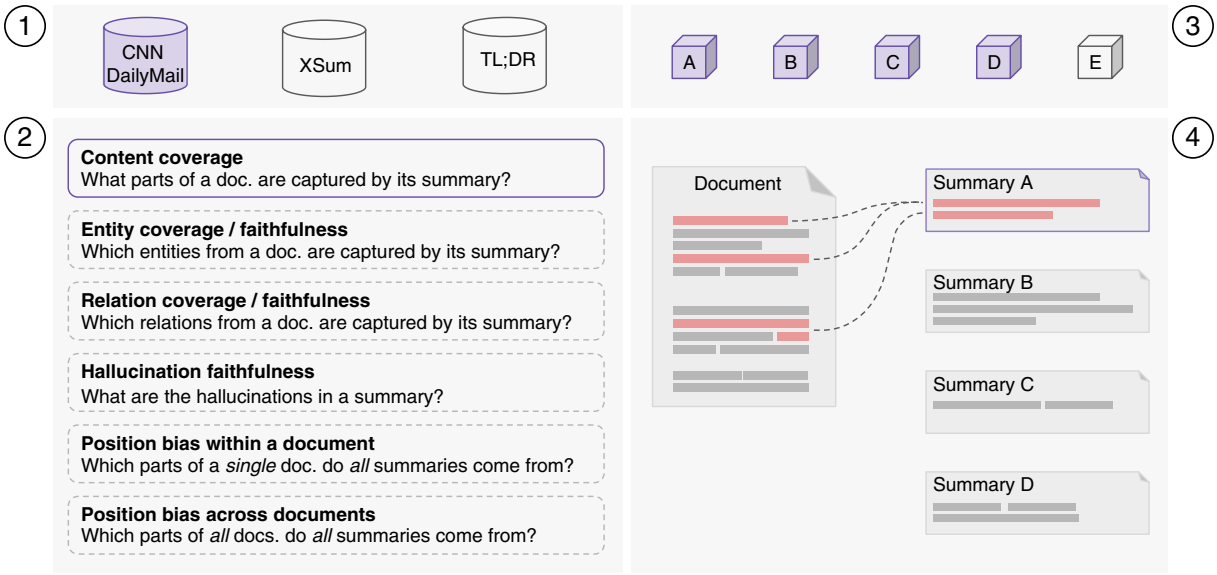

Figure 1: Overview of SUMMARY EXPLORER. Its guided assessment process works in four steps: (1) corpus selection, (2) quality aspect selection, (3) model selection, and (4) quality aspect assessment. Exemplified is the assessment of the content coverage of the summaries of four models for a source document from the CNN/DM corpus. For each summary sentence, its two most related source document sentences are highlighted on demand.

the state-of-the-art text summarization models, and to discover interesting cases that cannot likely be captured by automatic evaluation.

2 Related Work

Leaderboards such as Paperswithcode,[1] ExplainaBoard[2] and NLPProgress[3] provide an overview of state of the art in text summarization mainly according to ROUGE. These leaderboards simply aggregate the scores as reported by the models' developers, where the reported scores can be obtained using different implementations. Hence, a fair comparison become less feasible. For instance, the Bottom-Up model (Gehrmann et al., 2018) uses a different implementation of ROUGE,[4] compared to the BanditSum model (Dong et al., 2018).[5] Besides, for a qualitative comparison of the models, one needs to manually inspect the generated summaries, which are missing from such leaderboards.

To address these shortcomings, VisSeq (Wang et al., 2019) aids developers to locally compare their model's outputs with the ground truth, providing lexical and semantic comparisons along with statistics such as most frequent n-grams and sentence score distributions. LIT (Tenney et al., 2020) provides similar functionality for a broader range of NLP tasks, implementing a work-bench-style debugging of model behavior, including visualization of model attention, confusion matrices, and probability distributions. Closely related to our work is SummVis (Vig et al., 2021), the recently published tool that provides a visual text comparison of summaries with a reference summary as well as a source document, facilitating local debugging of hallucinations in the summaries.

SUMMARY EXPLORER draws from these developments and adds three missing features: (1) Quality-criteria-driven design. Based on a careful literature review of qualitative evaluation of summaries, we derive three key quality criteria and encode them explicitly in the interface of our tool. Other existing tools render these criteria implicit in their underlying design. (2) A step-by-step process for guided analysis. From the chosen quality criteria, we formulate concise and specific questions needed for a qualitative evaluation, and provide a tailored visualization for each question. While previous tools utilize visualization and enable users to (de)activate certain features, they oblige the users to figure out the process themselves, which can be overwhelming to non-experts. (3) Compilation of the state of the art. We collect the outputs of more than 50 models on three benchmark datasets providing a comprehensive overview of the progress in text summarization.

SUMMARY EXPLORER complements these tools and

[1] https://paperswithcode.com/task/text-summarization
[2] http://explainaboard.nlpedia.ai/leaderboard/task-summ/
[3] https://nlpprogress.com/english/summarization.html
[4] https://github.com/sebastianGehrmann/rouge-baselines
[5] https://github.com/pltrdy/rouge

also provides direct access to the state of the art in text summarization, encouraging rigorous analysis to support the development of novel models.

3 Designing Visual Summary Exploration

The design of SUMMARY EXPLORER derives from first principles, namely the three quality criteria *coverage*, *faithfulness*, and *position bias* of a summary in relation to its source document. These high-level criteria are frequently manually assessed throughout the literature. Since their definitions vary, however, we derive from each criterion a total of six specific aspects that are more straightforwardly operationalized in a visual exploration (see Figure 1, Step 2). To render the aspects more directly accessible to users, each is "clarified" by a guiding question that can be answered by a tailored visualization. Below, the three quality criteria are discussed, followed by the visual design.

3.1 Summary Quality Criteria

Coverage A primary goal of a summary is to capture the important information from its source document. Accordingly, a standard practice in summary evaluation is to assess its coverage of the key content (Paice, 1990; Mani, 2001; Jones et al., 1999). In many cases, a comparison to the ground truth (reference) summary can be seen as a proxy for coverage, which is essentially the core idea of ROUGE. However, since it is hard to establish an ideal reference summary (Mani et al., 1999), a comparison against the source document is more meaningful. Although an automatic comparison against it is feasible (Louis and Nenkova, 2013; ShafieiBavani et al., 2018), deciding what is *important* content is highly subjective (Peyrard, 2019). Therefore, authors resort to a manual comparison instead (Hardy et al., 2019). We operationalize coverage assessment by visualizing a document's overlap in terms of content, entities, and entity relations with its summary. Content coverage refers to whether a summary condenses information from all important parts of a document, measured by common similarity measures; entity coverage contrasts the sets of named entities identified in both summary and document; and relation coverage does the same, but for extracted entity relations.

Faithfulness A more recent criterion that gained prominence especially in relation to neural summarization is the faithfulness of a summary to its source document (Cao et al., 2018; Maynez et al., 2020). Whereas coverage asks if the document is sufficiently reflected in the summary, faithfulness asks the reverse, namely if the summary adds something new, questioning its appropriateness. Due to their autoregressive nature, neural summarization models have the unique property to "hallucinate" new content (Kryscinski et al., 2020; Zhao et al., 2020). This is what enables abstractive summarization, but also bears the risk of generating content in a summary that is unrelated to the source document. The only acceptable hallucinated content in a summary must be textually entailed by its source document, which renders an automatic assessment challenging (Falke et al., 2019; Durmus et al., 2020). We operationalize faithfulness assessment by visualizing previously unseen words in a summary in context, aligned with the best-matching sentences of its source document.

Position bias Data-driven approaches, such as neural summarization models, can be biased by the domain of their training data and learn to exploit common patterns. For example, news articles are typically structured according to an "inverted pyramid," where the most important information is given in the first few sentences (PurdueOWL, 2019), and which models learn to exploit (Wasson, 1998; Kedzie et al., 2018). Non-news texts, such as social media posts, however, do not adopt this structure and thus require an unbiased consideration to obtain proper summaries (Syed et al., 2019). We operationalize position bias assessment by visualizing the parts of a document that are the source of its summary's sentences, as well as the ones that are common among a set of summaries.

3.2 Visual Design

Guided Assessment SUMMARY EXPLORER implements a streamlined process to guide summary quality assessment, consisting of four steps (see Figure 1). (1) A benchmark dataset is selected. (2) A list of available summary quality aspects is offered each with a preview of its tailored visualization and its interactive use. (3) Applying Shneiderman's (1996) well-known Visual Information-seeking Mantra ("overview first, zoom and filter, then details-on-demand"), an overview of all models as a heatmap over averages of several quantitative metrics is shown (Figure 2a), which enables a targeted filtering of the models based on their quantitative performance. The heatmap of average values paints only a rough picture; upon model

Figure 2: (a) Heatmap overview of 45 models for the CNN/DM corpus; ones selected for analysis are highlighted red. Views for (b) the content coverage, (c) the entity coverage, (d) the relation coverage, (e) the position bias across models for a single document, (f) the position bias of a model across all documents as per lexical and semantic alignment, (g) the distribution of quantitative metric scores for a model.

selection, histograms of each model's score distribution for each metric are available. (4) After models have been selected, the user is forwarded to the corresponding quality aspect's view.

The visualizations for the individual aspects of the three quality criteria share the property that two texts need to be visually aligned with one another.[6] Despite this commonality, we abstain from creating a single-view visualization "stuffed" with alternative options. We rather adopt a minimalistic design for the assessment of individual quality aspects.

Coverage View (Figure 2b,c,d) Content coverage is visualized as alignment of summary sentences and document sentences at the semantic and lexical level in a full-text side-by-side view. Colorization indicates different types of alignments. For entity coverage (relation coverage), a corresponding side-by-side view lists named entities (relations) in a summary and aligns them with named entities (relations) in its source document. For unaligned relations, corresponding document sentences can be retrieved.

Faithfulness View (Figure 3, Case A) Hallucinations are visualized by highlighting novel words in a summary. For each summary sentence with a hallucination, semantically and lexically similar document sentences are highlighted on demand. Since named entities and thus also entity relations form a subset of hallucinated words, the above coverage views do the same. Also, in an aggregated view, hallucinations found in multiple summaries are ordered by frequency, allowing to inspect a particular model with respect to types of hallucinations.

Position Bias View (Figure 2e,f) Position bias is visualized for all models given a source document, and for a specific model with respect to all its summaries in a corpus. The former is visualized as a text heatmap, where a gradient color indicates for every sentence in a source document how many different summaries contain a semantically or lexically corresponding sentence. The latter is visualized by a different kind of heatmap for 50 randomly selected model summaries, where each summary is projected on a single horizontal bar representing the source document. Bar length reflects document length in sentences and aligned sentences are colored to reflect lexical or semantic alignment.

[6] A visualization paradigm recently surveyed by Yousef and Jänicke (2021).

Aggregation Options Most of the above visualizations show individual pairs of source documents and a summary. This enables the close inspection of a given summary, and thus the manual assessment of a model by sequentially inspecting a number of summaries for different source documents generated by the same model. For these views, the visualizations also support displaying a number of summaries from different models for a relative assessment of their summaries.

4 Collection of Model Outputs

We collected the outputs of 55 summarization approaches on the test sets of three benchmark datasets for the task of single document summarization: CNN/DM, XSum and Webis-TLDR-17. Each dataset has a different style of ground truth summaries, ranging from semi-extractive to highly abstractive, providing a diverse selection of models. Outputs were obtained from NLPProgress, meta-evaluations such as SummEval (Fabbri et al., 2021), REALSumm (Bhandari et al., 2020), and in correspondence with the model's developers.[7]

4.1 Summarization Corpora

The most popular dataset, CNN/DM (Hermann et al., 2015; Nallapati et al., 2016), contains news articles with multi-sentence summaries that are mostly extractive in nature (Kryscinski et al., 2019; Bommasani and Cardie, 2020). We obtained the outputs from 45 models. While the original test split of the dataset contained 11,493 articles, we discarded ones that were not summarized by all models, resulting in 11,448 articles total. This minor discrepancy is due to inconsistent usage by authors, such as reshuffling the order of examples, de-duplication of articles in the test set, choice of tokenization, text capitalization, and truncation.

For the XSum dataset (Narayan et al., 2018), the outputs of six models for its test split (10,360 articles) were obtained. XSum contains news articles with more abstractive single-sentence summaries compared to CNN/DM. The Webis-TLDR-17 dataset (Völske et al., 2017) contains highly abstractive, self-authored (single to multi-sentence) summaries of Reddit posts, although slightly noisier than the other datasets (Bommasani and Cardie, 2020). We obtained the outputs from the four submissions of the TL;DR challenge (Syed et al., 2019) for 250 posts.

[7] We sincerely thank all the developers for their efforts to reproduce and share their models' outputs with us.

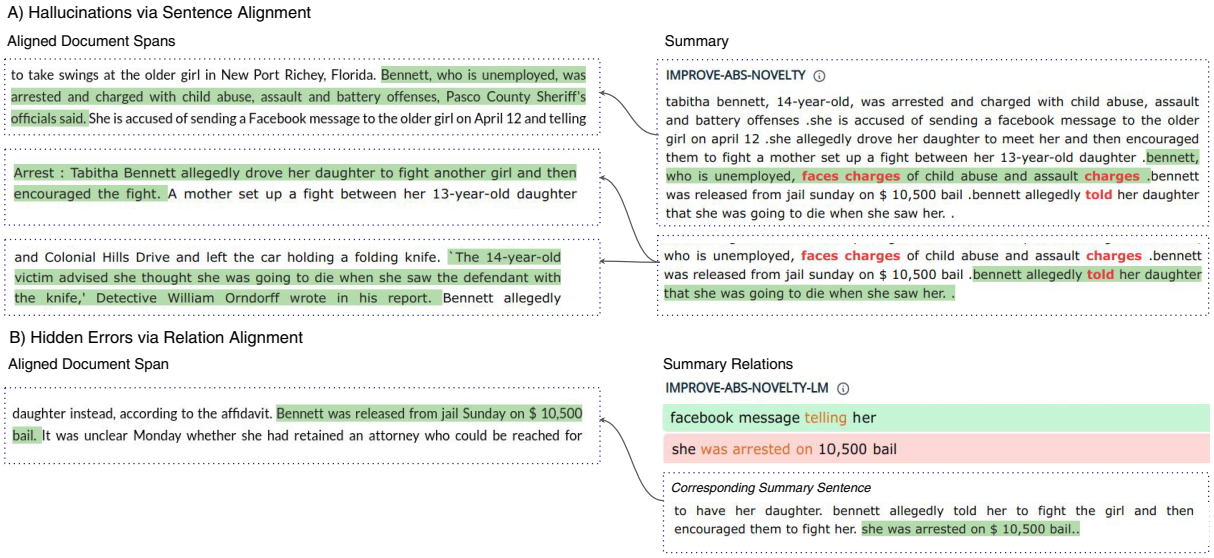

Figure 3: Two showcases for identifying inconsistencies in abstractive summaries using SUMMARY EXPLORER. Case A depicts the verification of the correctness of hallucinations by aligning document sentences. Case B depicts uncovering more subtle hallucination errors by comparing unaligned relations.

4.2 Text Preprocessing

In a preprocessing pipeline, the input of a collection of documents, their ground truth summaries, and the generated summaries from a given model were normalized. First, basic normalization, such as de-tokenization, unifying model-specific sentence delimiters, and sentence segmentation were carried out. Second, additional information, such as named entities and relations were extracted using Spacy[8] and Stanford OpenIE (Angeli et al., 2015), respectively. The latter extracts redundant relations where partial components such as either the subject or the object are already captured by longer counterparts. Such "contained" relations are merged into unique representative relations for each subject.

Alignment Every output summary is aligned with its source document, identifying the top two lexically and semantically related document sentences for each summary sentence. Lexical alignment relies on averaged ROUGE-{1,2,L} scores among the document and summary sentences. The highest scoring document sentence is taken as the first match. The second match is identified by removing all content words from the summary sentence already captured by the first match, and repeating the process as per Lebanoff et al. (2019). For semantic alignment, the rescaled BERTScore (Zhang et al., 2020) is computed between a summary sentence and all source document sentences, with the top-scoring two sentences as candidates.

[8]https://spacy.io

Summary Evaluation Measures Several standard evaluation measures enable quantitative comparisons and filtering of models for detailed analysis: (1) *compression* as the word ratio between a document and its summary (Grusky et al., 2018), (2) *n-gram abstractiveness* as per Gehrmann et al. (2019) calculates a normalized score for novelty by tracking parts of a summary that are already among the n-grams it has in common with its document, (3) *summary length* as word count (not tokens), (4) *entity-level factuality* as per (Nan et al., 2021) as percentage of named entities in a summary found in its source document, and (5) *relation-level factuality* as percentage of relations in a summary found in its source document. Finally, for consistency, we recompute ROUGE-{1,2,L}[9] for all the models.

5 Assessment Case Studies

We showcase the use and effectiveness of SUMMARY EXPLORER by investigating two models (IMPROVE-ABS-NOVELTY, and IMPROVE-ABS-NOVELTY-LM) from Kryscinski et al. (2018) that improve the abstraction in summaries by including more novel phrases. We investigate the correctness of their hallucinations (novel words in the summary), and identify hidden errors introduced by the sentence fusion of the abstractive models.

[9]https://github.com/google-research/google-research/tree/master/rouge

Hallucinations via Sentence Alignment Hallucinations are novel words or phrases in a summary that warrant further inspection. Accordingly, our tool highlights them (Figure 3, Case A), directing the user to the respective candidate summary sentences whose related document sentences can be seen on demand. For IMPROVE-ABS-NOVELTY, we see that the first candidate improves abstraction via paraphrasing, is concisely written, and correctly substitutes the term *"offenses"* with the novel word *"charges"*. The second candidate also improves abstraction via sentence fusion, where two pieces of information are combined: *"bennett allegedly drove her daughter"*, and *"victim advised she thought she was going to die"*. The novel word *"told"* also fits. However, the sentence fusion creates a wrong relation between the different actors (*"bennett allegedly told her daughter that she was going to die"*), which can be easily identified via the visual sentence alignment provided.

Hidden Errors via Relation Alignment The above showcase does not capture all hallucinations. SUMMARY EXPLORER also aligns relations extracted from a summary and its source document to identify novel relations. For IMPROVE-ABS-NOVELTY-LM, we see that the relation *"she was arrested"* is unaligned to any relation in the source document (Figure 3, Case B). Aligning the summary sentence to the document, we note that it is unfaithful to the source despite avoiding hallucinations (*"Bennett was released on $10,500 bail"*, and not *"arrested on $10,500 bail"*). The word *"arrested"* was simply extracted from the document sentence (Figure 3, Case A). Without the visual support, identifying this small but important mistake would have been more cognitively demanding for an assessor.

6 Conclusion

In this paper, we present SUMMARY EXPLORER, an online interactive visualization tool to assess the state of the art in text summarization in a guided fashion. In enables analysis akin to close and distant reading in particular facilitating the challenging inspection of hallucinations by abstractive summarization models. The tool is available open source[10] enabling local use. We also welcome submissions of summaries from newer models trained on the existing datasets as part of our collaboration with the summarization community. We aim to expand the tool's features in future work, exploring novel visual comparisons of documents to their summaries for more reliable qualitative assessments of summary quality. Finally, it is important to note that the accuracy of some of the views is influenced by the intrinsic drawbacks of the toolkits used for named entity recognition and information extraction.

[10] https://github.com/webis-de/summary-explorer

7 Ethical Statement

Visualization plays a major role in the usage and accessibility of our tool. In this regard, to accommodate for color blindness, we primarily use gradient-based visuals for key modules such as model selection, aggregating important content, and text alignment. This renders the tool usable also in a monochromatic setting. Regarding the hosted summarization models, the key goal is to allow a wider audience comprising of model developers, the end users, and practitioners to openly compare and assess the strengths, limitations and possible ethical biases of these systems. Here, our tool supports making informed decisions about the suitability of certain models to the downstream applications.

Acknowledgments

We thank the reviewers for their valuable feedback. This work was supported by the German Federal Ministry of Education and Research (BMBF, 01/S18026A-F) by funding the competence center for Big Data and AI (ScaDS.AI Dresden/Leipzig).

References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 344–354. The Association for Computer Linguistics.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages

8075–8096, Online. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4784–4791. AAAI Press.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. Banditsum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3739–3748. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5055–5070. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2214–2220. Association for Computational Linguistics.

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, Brussels, Belgium, October 31 - November 4, 2018*, pages 4098–4109. Association for Computational Linguistics.

Sebastian Gehrmann, Zachary M. Ziegler, and Alexander M. Rush. 2019. Generating abstractive summaries with finetuned language models. In *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*, pages 516–522. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 708–719. Association for Computational Linguistics.

Hardy, Shashi Narayan, and Andreas Vlachos. 2019. Highres: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3381–3392. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

K Sparck Jones et al. 1999. Automatic summarizing: factors and directions. *Advances in automatic text summarization*, pages 1–12.

Chris Kedzie, Kathleen R. McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1818–1828. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 540–551. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9332–9346. Association for Computational Linguistics.

Wojciech Kryscinski, 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 1808–1817. Association for Computational Linguistics.

Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Scoring sentence singletons and pairs

for abstractive summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2175–2189. Association for Computational Linguistics.

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

Elena Lloret, Laura Plaza, and Ahmet Aker. 2018. The challenging task of summary evaluation: an overview. *Lang. Resour. Evaluation*, 52(1):101–148.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Comput. Linguistics*, 39(2):267–300.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Reseach and Devopment*, 2(2):159–165.

Inderjeet Mani. 2001. Summarization evaluation: An overview.

Inderjeet Mani, David House, Gary Klein, Lynette Hirschman, Therese Firmin, and Beth Sundheim. 1999. The tipster summac text summarization evaluation. In *EACL 1999, 9th Conference of the European Chapter of the Association for Computational Linguistics, June 8-12, 1999, University of Bergen, Bergen, Norway*, pages 77–85. The Association for Computer Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 280–290. ACL.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cícero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathy McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2727–2733. Association for Computational Linguistics.

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, Brussels, Belgium, October 31 - November 4, 2018*, pages 1797–1807. Association for Computational Linguistics.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1925–1930. The Association for Computational Linguistics.

Chris D Paice. 1990. Constructing literature abstracts by computer: techniques and prospects. *Information Processing & Management*, 26(1):171–186.

Maxime Peyrard. 2019. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1059–1073. Association for Computational Linguistics.

PurdueOWL. 2019. Journalism and journalistic writing: The inverted pyramid structure.

Gerard Salton, James Allan, Chris Buckley, and Amit Singhal. 1994. Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, 264(5164):1421–1426.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond K. Wong, and Fang Chen. 2018. Summarization evaluation in the absence of human model summaries using the compositionality of word embeddings. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 905–914. Association for Computational Linguistics.

B. Shneiderman. 1996. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343.

Shahbaz Syed, Michael Völske, Nedim Lipka, Benno Stein, Hinrich Schütze, and Martin Potthast. 2019. Towards summarization for social media - results of the tl;dr challenge. In *Proceedings of the 12th International Conference on Natural Language Generation, INLG 2019, Tokyo, Japan, October 29 - November 1, 2019*, pages 523–528. Association for Computational Linguistics.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings*

of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020, pages 107–118. Association for Computational Linguistics.

Jesse Vig, Wojciech Kryscinski, Karan Goel, and Nazneen Fatema Rajani. 2021. Summvis: Interactive visual analysis of models, data, and evaluation for text summarization. *CoRR*, abs/2104.07605.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. Tl;dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization, NFiS@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, pages 59–63. Association for Computational Linguistics.

Changhan Wang, Anirudh Jain, Danlu Chen, and Jiatao Gu. 2019. Vizseq: a visual analysis toolkit for text generation tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019 - System Demonstrations*, pages 253–258. Association for Computational Linguistics.

Mark Wasson. 1998. Using leading text for news summaries: Evaluation results and implications for commercial summarization applications. In *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th Int. Conference on Computational Linguistics*, pages 1364–1368.

Tariq Yousef and Stefan Jänicke. 2021. A survey of text alignment visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1149–1159.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online. Association for Computational Linguistics.